DOCUMENT RESUME

ED 229 358 SP 022 233

AUTHOR	Denton, Cliff; Postlethwaite, Keith
TITLE	The Identification of More Able Pupils in Comprehensive Schools. Final Report. Part I.
INSTITUTION	Oxford Univ. (England). Oxford Educational Research Group.
PUB DATE	Dec 82
NOTE	213p.; For related document, see SP 022 234.
PUB TYPE	Information Analyses (070) -- Reports - Research/Technical (143)
EDRS PRICE	MF01/PC09 Plus Postage.
DESCRIPTORS	*Ability Identification; *Academically Gifted; *Checklists; Foreign Countries; *Informal Assessment; Secondary Education; Secondary School Teachers; Skill Analysis; Standardized Tests; Student Characteristics; Student Evaluation; *Teacher Attitudes; *Test Reliability
IDENTIFIERS	England

ABSTRACT

A 2-year project investigated the ability of secondary school teachers to identify the top 10 percent of students in each of four subjects, English, French, physics, and mathematics. The students were tested on a wide range of aptitudes and attitudes at the beginning of their third form (ninth grade). Teachers were asked to list students they considered high or outstanding in potential in their subject. Encouraging levels of agreement were found in all four subjects, though there was some evidence of errors in physics and French. Test-identifiable characteristics of the teacher-selected students were examined to discover clues to possible biases affecting teacher judgment. Physics teachers tended to place emphasis on general ability rather than subject-specific abilities, while French teachers tended to select students with favorable attitudes toward the subject. Subsequently, the school sample was divided into two groups. One group of teachers made ability selections in the same way as before, while the other filled in a subject-specific profile of aptitudes using checklists as an aid to identification. The conclusion was reached that the checklists had little effect on the ability of teachers to identify able students. (JD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED229358

SP

OXFORD EDUCATIONAL RESEARCH GROUP

THE IDENTIFICATION OF MORE ABLE PUPILS IN COMPREHENSIVE SCHOOLS

by

Cliff Denton

and

Keith Postlethwaite

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY
F. C. J. Denton

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

FINAL REPORT
PART I

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Presented to the
Department of Education and Science
December 1982.

SP022 233

The Identification of More Able Pupils
In Comprehensive Schools

PART I

NOTE: For the convenience of binding the report has been printed in two parts. Part I deals with the research design and the work of the first year. Part II deals with the work of the second year and includes a summary of the conclusions and also the appendices for the whole report.

CONTENTS

PAGE

ABSTRACT i
ACKNOWLEDGEMENTS iv
STEERING GROUP vi

PART I

CHAPTER ONE - Defining Basic Questions 1
- 1.1 Introduction 1
- 1.2 Findings from visits made 6
- 1.3 A review of the relevant literature 9
- 1.4 Summary 21

CHAPTER TWO - Research Design 22
- 2.1 Introduction 22
- 2.2 Outline of the research design 22

CHAPTER THREE - The Agreement Between Test and Teacher Nomination of More Able Pupils 38
- 3.1 Introduction 38
- 3.2 Brief summary of the conclusions 39
- 3.3 Preliminary investigation of the match between test and teacher based identification 40
- 3.4 Comparison of teacher nomination with test prediction when test error is taken into account 48
- 3.5 The match between teacher and test judgements in terms of Groups A, B and C 52
- 3.6 The level of mismatch in Groups A and C 57
- 3.7 The mismatch between test and teacher identification when fuller allowance is made for the test error 59
- 3.8 Subsets of the top 10% 68

CHAPTER FOUR - A Preliminary Investigation of the Characteristics of Children who were Nominated as More Able by their Teachers 73
- 4.1 Introduction 73
- 4.2 Summary of conclusions 75
- 4.3 Characteristics of the teacher selected group compared with the whole of the rest of the year group 79

PAGE

CHAPTER FIVE - Preliminary Work on Checklist-Aided Identification 128

5.1 Introduction 128

5.2 Research method for the preliminary study of checklists 130

5.3 Construction of the checklists 136

5.4 Method of use 147

CHAPTER SIX - The Effect of the First Version of the Checklist 149

6.1 Introduction 149

6.2 Summary of the results 149

6.3 Level of agreement between test prediction and teacher judgement 151

6.4 A comparison of results in checklist and non-checklist schools in Round 2 and of the results of each group of school with the results of Round 1 156

6.5 An investigation concerning the test scores of pupils who were nominated in one round but not the other 160

6.6 Test based characteristics of pupils chosen in Round 2 164

CHAPTER SEVEN - The Interview Study of Pupils About whom Tests and Teachers Disagreed 173

7.1 Introduction 173

7.2 Outcome 174

7.3 Results for English 176

7.4 Results for French 180

7.5 Results for Physics 184

7.6 Results for Mathematics 190

7.7 General conclusions 195

CHAPTER EIGHT - Summary of Findings from the First Year 196

8.1 Introduction 196

8.2 Summary 196

PAGE

PART II

CHAPTER NINE - The Second Year 201

CHAPTER TEN - An Approach to the Understanding of Teachers' Views of Pupils Based on Personal Construct Theory 203

CHAPTER ELEVEN - Results of Personal Construct Study 218

- 11.1 Basic results in Physics 218
- 11.2 Preliminary conclusions in Physics 233
- 11.3 Further analysis in Physics 233
- 11.4 Basic results in English 243
- 11.5 General conclusions for English 251
- 11.6 Basic findings in French 252
- 11.7 General conclusions in French 257

CHAPTER TWELVE - Further Development and Use of Checklists 261

- 12.1 Introduction 261
- 12.2 Development of further use of checklists in English, French and Physics 264
- 12.3 Mathematics 271

CHAPTER THIRTEEN - Conclusions and Recommendations 328

- 13.1 Introduction 328
- 13.2 Summary of general conclusions 328
- 13.3 Identification by test or by teacher 332
- 13.4 Use of checklists 335
- 13.5 A second observer 335
- 13.6 Identification through provision for a larger group 336
- 13.7 In-service training 337
- 13.8 Pupil grouping 338

PAGE

APPENDICES

1. Test battery and prediction equations 340
2. Forms used by teachers for listing top 10% of pupils 355
3. Further details concerning the school pupil sample 359
4. Further statistical tables 362
5. Instructions to teachers and checklists that were used in Years 1 and 2 that have not been included elsewhere in the report 382
6. Concerning the sample from which the teacher nominated the top 10% 405
7. The Mathematics checklist (Validity and Reliability) 407
8. Teachers constructs in Mathematics 477

BIBLIOGRAPHY 484

ABSTRACT

The Identification of More Able Pupils in Comprehensive Schools

The project was concerned with the ability of teachers to identify the top ten per cent and smaller subgroups of children in each of four subjects, English, French, physics and mathematics.

The research fell into two distinct phases of activity each of which took up one year of the two year life of the project.

In each of the two phases, eleven schools which comprised a representative sample of comprehensive schools in the County of Oxfordshire co-operated in the investigation. Two successive third forms of pupils from these schools were the subjects of the study (2400 in the first year and 1500 in the second).

These pupils were tested on a wide range of aptitudes and attitudes at the beginning of their third form. Subject specific equations were derived by multiple regression analysis which enabled predicted O level scores to be calculated for each of the pupils in each of the four subjects. The equations themselves were derived from data concerning a completely different set of pupils in two schools who had undergone the same tests and also gone on to take their O level examinations. The prediction for O level by this means of combining appropriate test scores was more accurate than would have been possible by using a simple I.Q. test. An estimate of the average error of prediction was also made.

In the first phase of the research, teachers were asked to list the pupils who they considered to be in the top 10% with regard to potential in their subject and also to list the top 5% and any outstanding pupils. A large scale statistical exercise was then mounted in order to assess the agreement between test and teacher.

Encouraging levels of agreement were found in all four subjects, though in physics and French there was enough evidence of errors on the part of teachers to cause some concern.

The test identifiable characteristics of the teacher selected pupils were examined in order to discover clues to the kinds of biases that might be affecting teacher judgement. Biases certainly existed but were different in nature from teacher to teacher. These differences did not seem to be particularly dependant on thetype of school in which the teacher worked. A considerable number of physics teachers, seemed to place undue emphasis on the general ability of pupils rather than on more appropriate subject specific abilities. In French there was evidence that pupils who had favourable attitudes to the subject tended to be selected.

Later in the year the school sample was randomly divided into two groups. One group of teachers acted as control and were asked to make a second identification in the same way as before while the other group of teachers were asked to fill in a subject specific profile of aptitudes for each child and use these 'checklists' as an aid to the second identification.

In all schools the level of agreement between test predicted scores and teacher nomination was very close to that measured from the first round, though there was some change in the nature of the pupil characteristics that were associated with teacher selection. In particular, the extra time that physics teachers had known their pupils led them to a large scale revision of their lists and though the agreement with test scores was no better, the bias towards general ability was not strongly evident. The other subjects showed little change and, in particular, the concerns regarding the characteristics of pupils nominated in French remained.

The main conclusion, however, was that checklists, suprisingly, had little effect on the ability of teachers to identify able pupils.

The second phase of the research sought answers that the statistical work could not find and which were difficult to untangle from the interviews that were held with teachers concerning their strategies.

First a method of analysis based on Personal Construct Theory was developed to attempt to isolate the kinds of characteristics that teachers observed in their pupils and to understand the different ways in which these characteristics influenced teachers in carrying out the task of identification.

It was discovered that the components of most teachers' knowledge of individual pupils was not very extensive. In particular (except in English)

the teachers' maps of pupils' subject specific abilities were sparse.

There were a number of behavioural and motivational influences on teacher judgements as well as those based on pupils test results. We were able to achieve considerable understanding of this aspect of the identification process as a result of this work.

Secondly we attempted to find ways of improving the method of use of the checklists and to come to an understanding of why they would be ineffective in their normal use. In one subject (mathematics) extensive classroom observations provided evidence that checklists could be an instrument upon which complete reliance might be placed. It was possible for an observer to classify pupils on the checklist items, according to the pupils' reactions to the challenges within the classroom, and to then make valid identifications based on the records. It was, nevertheless, not possible to do this without deliberately making use of time, when appropriate, to actively seek clues. The method was to work with individual children and modify the sorts of challenge presented to the pupils so as to take in the whole spectrum of aptitudes that are appropriate to the subject. Considerable understanding was achieved concerning the inability of a teacher, pressurised by time and by syllabus to identify by means of checklist.

The relationship between identification and provision was quite clearly demonstrated.

ACKNOWLEDGEMENTS

No worthwhile work of this kind can be done without the co-operation and support of many people. Our thanks go to all who have helped or encouraged us in any way. In particular we would like to offer thanks:

to Dr. Harry Judge whose unceasing support and confidence have been a source of strength, and without whom the work would not even have started;

to the administrators of the Hulme Surplus Fund at Brasenose College who provided funds for a one year secondment of one of the research officers so that a detailed research proposal could be developed;

to the large number of colleagues involved in related activities throughout the country, who gave freely of their valuable time in both the planning stage, and during the research phase;

to the officers of the Oxfordshire LEA who gave their support, not just in principle, but also by allowing a number of secondments, so that teachers could become actively involved in the project;

to the three teachers who, during the planning stage, were seconded part-time to the project; and to one in particular, Chris Davies, who subsequently joined the project for one day a week for a further two years to advise on English and French, and to do much of the work with the schools in these subjects;

to the headteachers of the schools involved with the project, for their help in organising our school based work and their time in discussing interim findings with us;

to the teachers who have helped us in so many aspects of the work when there were always many other demands on their time and energy;

to our colleagues at the University of Oxford Department of Educational Studies for their inspiration and advice, and to others in the University, especially at the Department of Experimental Psychology for their guidance on certain important aspects of the work;

to the Oxford Delegacy of Local Examinations and to The Oxford and Cambridge Schools Examination Board for the confidential release of examination scores necessary for the background work;

to Mrs. P. Athawes and Mr. J. Hailstone of the Rutherford Laboratory, Computing Division, for their help in the machine marking of part of our backgorund test battery;

to the Steering Group under the chairmanship of Divisional Inspector Mr. D.T.E. Marjoram for their valuable guidance in formal meetings and for their help in informal discussions at other times;

to Bob Hext of the University of Oxford Department of Educational Studies for the truly professional management of our administrative matters;

finally to our secretary Danielle Andersen, and to Sue Kitt whose perseverence with impossible handwriting has made the report readable for others.

It goes without saying, of course, that the shortcomings which remain in this report are entirely the responsibility of the authors.

CD/KCP
Oxford
December 1982

STEERING GROUP

The members of the Steering Group for the two years of the project were:

CHAIRMAN - Mr. D.T.E. Marjoram (Divisional Inspector)

Mr. W.M. Caldow (D.E.S.)
Dr. H.G. Judge (Director, University of Oxford Dept. of Educational Studies)
Mr. B.E. Day (Deputy Chief Education Officer of Oxfordshire)
Dr. E.M. Hitchfield (formerly of the University of Warwick)
Dr. B.E.F. Fender (Lecturer in Inorganic Chemistry, University of Oxford)
Mr. D.E.R. George (Dean of Sciences, Nene College, Northampton)
Mr. J.M. Deans (Deputy Head, Bartholomew School, Eynsham)
Miss A. Wood (Deputy Head, Milham Ford School, Oxford)
Research Officers

CHAPTER ONE

Defining Basic Questions

The purpose of this chapter is to show how the detailed research questions tackled in this study were defined through discussions with local teachers, through visits to existing projects which are working with able children and through a survey of the literature. Essentially the visits indicated that teachers felt a considerable need for detailed information about how to identify pupils who had high potential in particular academic subjects. A review of the literature indicated that this subject specific focus was sensible and provided some theoretical support for the view (which often finds expression in practice) that a major component in any identification scheme should be the professional judgement of the teachers.

The review revealed that little was known about teachers' ability to carry out this task of identification, although it also suggested that, in a closely related field (the identification of high IQ children) they were not very effective. As a result, there is often a suggestion that teachers need guidance in carrying out this task. Our visits showed that, in practice, this guidance is usually offered in the form of checklists on which teachers are asked to indicate whether a pupil displays certain behaviours which the checklist designer regards as indicative of high ability. Despite the extent to which this form of guidance is recommended, there is very little evidence concerning its effectiveness.

This background led us to define our research as a study of teacher-based identification of pupils with high potential in particular academic subjects and to take, as a major issue in this study, the question of the effectiveness of checklists in guiding teachers' judgements.

We will now develop these points in some detail.

1.1 Introduction

The Oxford Educational Research Group, under whose auspices this project was conducted, is a partnership between the Oxfordshire schools, the LEA, the University of Oxford Department of Educational Studies and other departments of the University. It aims to carry out research into matters which are of contemporary concern to those actively engaged in education and to do so in such a way that the work can lead to improvements in practice. All members of the partnership are encouraged to take an active part in all stages of the research - in defining the problem to be tackled, in steering the work once it has begun and in the day to day business of collecting data and discussing interim findings.

Because of these basic principles of OERG, the focus of our project was defined through an extensive series of discussions undertaken by one of the research officers in 1978/79. Visits were made to a large number of secondary schools in Oxfordshire and to other places in the country (such as university departments of education, teacher training colleges and LEA advisory services) where work on the gifted secondary school pupil was under way. These visits indicated a very real interest on the part of the majority of these practitioners in one particular, though still broad, problem within the field of what might be called 'gifted education'. That problem was how to provide an appropriate education for pupils who, relative to their peers in their own school, had high potential in a particular area of the academic secondary curriculum - pupils whom we, in shorthand, have called 'more able pupils' in the rest of this report.

Our visits to schools indicated that there was a very strong consensus of opinion amongst teachers on this issue. They felt that, as individual secondary school teachers, there was little that they could do to make general provision for the generally able pupil and that, in the course of their normal work with pupils, there was little opportunity to gain a wide enough view of pupils' abilities to enable them to make a confident identification of those who were 'generally able'. The mathematician, say, was interested in making appropriate provision for the able mathematics pupil. Whether these pupils were of high general ability, or of average (or even very occasionally remedial) ability in other subjects was of interest only in deciding how to make the mathematics provision work, and in determining how much priority should be given to it in planning the overall programme for the pupil. It was not important in confirming or denying the right of that pupil for membership of the more able mathematics group.

As well as drawing our attention to the importance teachers attached to this idea of a subject specific study of the more able, our visits to schools also emphasised that teachers wished attention to be focussed on quite a large group of pupils in any one subject area. The feeling was that far more than the top 0.1, 1 or even 2% who have traditionally been regarded as gifted, needed some form of additional, or alternative provision if their education was to be entirely appropriate to their ability.

Although the HMI survey of secondary education (HMI, 1979) was not exclusively concerned with more able pupils, it was interesting to see reflected in it, on a national scale, evidence of the same concerns about able children in specific subjects that had been expressed to us by the teachers whom we interviewed. For example HMI, talking of the need for greater stimulus for the more able in mathematics, stated that:

> Appropriate provision for these abler pupils (whether future students of A-level mathematics or not) could involve an extension of the ordinary syllabus, but the greater need relates to outlook, pace and variety of application. The pupils require a challenge to work in different ways; there should be more opportunity for extended investigations of topics which catch the imagination, and much more encouragement to undertake supplementary reading.

Again, writing of science, HMI commented:

> In about one third of all schools it was apparent that insufficient demands were made of able pupils. For example, those who finished practical work long before the rest of the class were frequently left with nothing to do when perhaps additional experimental work of a problem solving kind could have challenged them and tested their abilities to the full.

There was therefore no doubt that questions relating to more able pupils in a subject specific context could be seen as examples of contemporary concern at national, as well as local level.

This same HMI survey also provided support for the teachers' view that a relatively large group of pupils should be considered as more able. It is not easy to be clear from the report exactly the proportion of pupils to which HMI were referring when they wrote of the problems of the more able, but there is a strong implication that it was a fairly large group of perhaps 15% of the population in a school. What is quite clear is that HMI were not commenting on just the top 2% (for there is a separate Appendix to the report that deals with these pupils). Further independant support for the study of a fairly large group of pupils is provided by Renzulli and colleagues (1981) in their 'Revolving Door Model' of identification to which we will refer again later.

As we carried out our visits to schools it became clear that one particular question in the general area of concern about more able pupils in a given subject was repeatedly being brought to our attention. This was the question of how such pupils are to be identified for any additional or alternative activities which might be arranged in an attempt to make their educational diet more appropriate to their special needs. Again, there was support, from a national survey for the idea that identification did indeed pose a problem. In their report specifically about gifted children, HMI (1977) recommended that a systematic screening procedure should be evident in every school but reported that this did not generally exist. They commented that:

> There is no overall policy for identification. Identification is a hit and miss affair, the initiative coming mainly from individual teachers.

Their support for the idea that identification should be carried out on a subject specific basis can be seen in the fact that the second section of the report is divided into chapters which deal separately with different subjects. Most of these chapters make direct reference to how pupils, more able in that subject, could be recognised. The rich diversity of characteristics listed in this part of the report give great weight to the idea that ability in one subject may not necessarily go hand in hand with ability in another.

Thus our discussions with teachers, coupled with our reading of the HMI reports led us to the view that useful work of relevance to one of the contemporary concerns of practitioners in education could be done by studying matters relating to the identification of more able pupils in a subject specific context. They also emphasised that the target group for such a study should be fairly large. We will move on, in the next section to give further context for such a study which will be derived from both a survey of the relevant literature and from the findings of our other visits to more able pupil projects in Britain.

However before we embark upon this, it is interesting to draw attention to the fact that the opinions of teachers which we have been reporting above, give an indication of a very considerable shift in attitudes towards gifted children over the last decade. This change, to a situation in which teachers are more favourably disposed towards the idea of considering the more able as a group which does have special needs in education, is in itself important, as it implies that new information in the still under-researched areas of how, effectively, to identify and provide for these children, is likely to be seized by teachers and drawn into their future practice. The position in the early 1970s was recently described in an article by Ogilvie (1980):

> The position in schools in 1970 with regard to the concept of 'giftedness' almost defied description. Some teachers felt very strongly that the whole idea smacked of elitism, and hence were reluctant even to discuss it; others were equally certain that the only real difficulty lay in problems of reliability of intelligence tests.

Our discussions would indicate that this was no longer an accurate indication of the way in which our teachers thought about gifted children. They were no longer straight-jacketed by the concept of intelligence testing. They were also much more favourably disposed towards the general question of making provision for the gifted. Positive attitudes, similar to those which we met, were also reported by Tilsley (1979). The results of his survey implied that only about 10% of the teachers in his sample had neutral or negative attitudes towards the issue of the gifted pupil.

The reasons for the change in general attitude over ten years are no doubt complex and we will not attempt to trace them in detail here. One component may be that the commonly recommended technique of making provision in the form of enrichment materials raises fewer problems of elitism in the minds of teachers than did earlier suggestions for provision - which tended to lay emphasis on techniques such as acceleration or 'express streams'. These tended either to be limited in application to a smaller range of pupils, or to have significant resource implications. There would seem to be some support for the importance of the resource issue in determining attitudes in the detailed findings of Tilsley's study. Another component, however, seems to be that teachers are becoming increasingly interested in the importance of trying to meet an individual pupil's particular educational needs wherever that pupil lies on the spectrum of abilities. Work with the gifted is therefore often viewed as one aspect of a much wider activity aimed at meeting the whole range of special needs. It has, therefore become linked in the mind, and sometimes in location in the school - though not in the letter of the law - with work aimed at the very different groups of children who are the subjects of the 1981 Education Act. This, as we would suggest is a very proper link, which may well provide the environment in which the present positive attitudes towards the gifted will be encouraged to grow.

The reasons for the increased willingness to think in terms other than general intelligence are even more difficult to trace. Maybe the subject-based projects which are now well known (and on which we will report in Section 1.2 below) have had some impact on attitudes of a group of teachers much wider than that formed of those who are actively involved. Maybe the changing general attitudes have encouraged teachers to think about how they could help to provide for more able pupils - and this has driven them towards a subject specific model. On the other hand it may be the preparedness of teachers to accept a subject specific model that has encouraged them to move towards a more positive general attitude! Clearly an interesting, and important study could be done in untangling this web of cause and effect. We, however, will take it no further in this report.

1.2 Findings from visits made.

We have reported above that the HMI surveys (1977, 1979) themselves discovered, and described many shortcomings of schools in identifying more able pupils in a systematic way. We have also already shown that our own visits to schools confirmed that our teachers were very aware of this problem. We found that there was a strong sense of need for a well researched strategy for identification and that teachers felt a lack of confidence in their own ability to identify the more able without guidance.

Our visits to the major programmes for able children throughout the country were therefore carried out in order to see if there was already a well researched technique available, even if this had not yet found widespread application in schools. These visits revealed, however, that many of the projects tended to rely on teachers' judgements and that they did not undertake, as a major activity, any detailed evaluation of the method of identification that they used. The general feeling expressed was that this would have diverted energy from their prime, and entirely proper concern: namely that of making provision for children. Their starting point was typically that more able children had special educational needs that were not being fully met by the normal routine, and that the projects' efforts should be concentrated on developing ways of meeting these needs. This is not to say that the techniques that they adopted for identification were not based on a careful assessment of the relevant literature. It is only to say that nothing new could be learned of the effectiveness of the chosen identification strategies from the projects themselves.

It is perhaps helpful to illustrate the position with some references to particular projects. For example, the Nottinghamshire project, which was for a time one of the largest programmes for able children in the country, in its recommendations on the identification strategies to be used in its pilot study, assumed that teacher nomination would be the starting point of the screening process, and this would be followed up by input from the county's Psychological Services. The provision was to be subject specific, so that teacher nomination within a subject was called for. The *assumption* was that this would be effective.

The Teacher Education Project in the pamphlet '*Teaching Bright Children*' (Kerry 1981) set out guidelines for the identification process, broadening the definition of 'bright' from those who perform well at tests of general ability (IQ) to those who have this general ability and/or other high abilities of more specific kinds. An attempt was made to systematise the identification process for this more broadly defined group. A checklist approach to identification was recommended and examples of checklists, such as that put forward by Ogilvie (1973), were given. Despite an extensive set of guidelines, however, there was an *assumption* that an effective strategy, based on these guidelines could then be adopted for each teacher's own situation.

As in several other LEAs, the identification recommended by the Essex LEA, which has devoted considerable resources to the enrichment of able pupils, was along the lines recommended by Kerry and thus made the same assumptions.

Provision in science has been taken up in a number of parts of Britain. The Cleveland science checklist, now included in the Schools Council booklet on identification strategies (Clarke, 1981), was developed for the purpose of identifying pupils for that county's science programme for able children, but a scheme for monitoring its effectiveness was not mounted. Similarly, identification based upon teacher nomination was used in the Nene College programme and in the ESTEEM project at Warwick (Screen, 1982), again without evaluation.

These few examples illustrate what was found in general from the round of visits: namely that the methods of identification used in the major programmes for the more able pupil were firmly based on teachers' judgements and that no programme was designed so that it could give valid evidence of the effectiveness of its own chosen procedures. We repeat that this is not to imply criticism of these other projects; their aims were different, and their structure did not enable them to include a detailed research strand in their overall plan of campaign. The point is simply that a detailed evaluation of the common identification methods - and in particular, of checklist guided, teacher based identification - was now needed. What was also clear was that this evaluation could emerge from the research that we were planning.

As well as emphasising the need for such a piece of research, there was a further valuable consequence of the visits that we made. This was that they helped us to define more clearly the detailed questions that should be answered by our research programme. It became clear that these question should include:

As teachers are so often the main agents in identification for subject specific programmes, are they effective at identifying able pupils in a particular subject?

Does their effectiveness vary from subject to subject?

If teachers need guidance - as is suggested in several of the reports above - what is the nature of their difficulty, and therefore of the most appropriate guidance which can be provided?

Since a very common form of recommended guidance is based upon the use of checklists, does the use of an appropriate checklist actually improve teachers' effectiveness?

Since teacher-based identification is recommended as a way of selecting pupils for provision which is organised on an out of school basis, is the effectiveness of teacher-based identification affected by the type of school in which the teacher operates?

Our next task was to explore the literature in an attempt to determine whether the suggested subject specific bias for identification was sensible, to discover whether there was any positive reason for the common view that identification should be based on teacher judgement, to find out what was already known about the accuracy of teachers in identifying pupils of high ability in specific subjects and to assess the value of checklists in guiding this process. It is to this review of the literature that we now turn.

1.3 A review of the relevant literature

In their pioneering work on the gifted, Terman and his associates (1925, 1926, 1930, 1949, 1959) took as their main means of identification, pupils' scores on the Stanford-Binet Intelligence Scale. Pupils scoring 140 or more on this scale were included in their gifted group. In their subsequent study of the characteristics of such pupils, Terman's team discovered that they were:

> appreciably superior to unselected (i.e. IQ less than 140) children in physique, health and social adjustment, markedly superior in moral attitudes.... and vastly superior in their mastery of school subjects.

They also established the fallacy of the view encapsulated in the phrase 'Early ripe, early rot'. They thus overthrew many misconceptions about the gifted which found expression in such writings as 'The Insanity of Genius' (Nisbet, 1895).

It was natural that a massive innovatory study such as Terman's should generate much activity, and with a well developed test that could be used to identify children of high IQ, important research into the characteristics of this kind of gifted child flourished. Notable contributions to work in this vein were provided by, among others, Hollingworth (1942), Parkyn (1948), Burt (1962), Lovell and Shields (1967). There was considerable agreement between these researchers as to the characteristics of this kind of gifted child and in consequence the picture that their work painted of such children has become widely accepted - and perhaps too widely applied.

As this work continued, a certain amount of critical interest was expressed in the fact that there was such great emphasis placed on the single criterion of high IQ. One aspect of this concern rested upon the fact that the cut-off point for a pupil to be included in the gifted group was arbitrarily defined. However an even more serious concern was expressed by workers who felt that the emphasis on the single dimension of IQ was unhelpfully restrictive. Witty (1951) suggested that we should move away from an IQ based definition of giftedness towards a much broader view. He offered, as an alternative, the suggestion that:

> the definition of the gifted be expanded to include any child whose performance in a worthwhile type of human endeavour is consistently or repeatedly remarkable.

Later, Getzels and Jackson (1958) continued the attack against undue emphasis on IQ by writing that:

> it not only restricts our perspective of the more general phenomenon, but places on the one concept a greater theoretical and predictive burden than it was intended to carry.

Other more recent writers have also attempted develop this philosophy. In an influential report to Congress, the U.S. Commissioner for Education gave the following definition of the gifted (Marland, 1972):

> (They have) high demonstrated achievement, and/or potential in any of the following areas, singly or in combination:
>
> general intellectual ability;
> specific academic aptitude;
> creative or productive thinking;
> leadership ability;
> visual and performing arts;
> psychomotor ability.

In Britain, a little later, Ogilvie (1973) suggested the following definition of the gifted pupil:

> The term 'gifted' is used to indicate any child who is outstanding in either a general or specific ability, in a relatively broad or narrow field of endeavour ... Where generally recognized tests exist as (say) in the case of 'intelligence', then 'giftedness' would be defined by test scores. Where no recognized tests exist it can be assumed that the subjective opinions of 'experts' in the various fields on the creative qualities of originality and imagination displayed would be the criteria we have in mind.

And more recently still, Kerry (1981) wrote:

> We shall call 'bright' any child with IQ in the region of 130 or one who shows an outstanding talent in any one field of schoolwork which sets him or her notably above other pupils in the class or age-group.

Such definitions suggest that the subject specific focus that we were urged to adopt for our own study was a reasonable one, in that pupils with high potential in a specific subject would certainly fall within the 'gifted groups' proposed by these definitions. What the definitions also did of course, was to warn us, as they will undoubtedly warn our readers, that our study could not be regarded as a complete study of all aspects of the identification of the gifted. Some high specific abilities (e.g. an exceptionally developed moral sense or unusually high ability as a leader) that would certainly fall within the scope of the definitions would not fall under any subject specific heading whichever subjects were included in our work. What our work does then is to explore identification in relation to one part of the definition. In doing so it breaks new ground in that earlier work tends to be focussed on another part - namely that relating to high IQ. We will return to this point later but next will explore something of the theoretical background to the definitions themselves.

Important theoretical support for the evolution of these broader concepts of giftedness was to be found in studies which attempted to explore the structure of human ability and to isolate its separate components. Such attempts have a long history. In the light of the work of Torrance (1961) and Getzels and Jackson (1962) the idea grew that 'creativity' was one such component. (In the context of our interest in how to identify the 'gifted' it is important to note that the latter researchers indicated not only that many highly creative pupils would be overlooked by IQ tests but also that such pupils tended to be unpopular with teachers - and so might easily be overlooked if identification was to be based on teachers' judgements).

However, important though the concept of creativity is, it is not the only additional component of ability that began to make inroads on the central role of IQ in the definition of giftedness. Thurstone's (1941) suggestion that the IQ test itself was a composite of many cognitive tests raised the possibility that a more systematic mapping of the cognitive domain might bring to light several quite specific forms of cognitive ability. Guilford (1959, 1967) provides the most extreme example of the outcome of such a mapping. He suggests that the intellect is made up of 120 separate specific abilities, and has produced tests which seem to measure many of them in isolation. His view has been severely criticised. Eysenck (1961) suggests that it is a reductio ad absurdum, that the statistical technique is capable of infinite subdivision, that the model is a poor predictor and that it omits the essentially hierarchical nature of the data. Indeed Guilford himself admits that the details of his model 'may or may not stand the test of time' though he does go on to say that the underlying principle of the 'multiplicity of intellectual abilities seems well established' (Guilford in Barbe and Renzulli, 1975). The exact relationship, then, between general ability (as implied by the adoption of an IQ model) and specific independant abilities (as suggested in the Guilford model) remains a matter of debate and it seems that until further work has been done it would be wise to keep both models in mind.

Vernon (1969) has provided an interesting analysis of how the two models can be used without undue conflict. He suggests that:

> when dealing with homogeneous or highly selected populations it is profitable to study different mental faculties and abilities along special lines.......but with more heterogeneous populations, and particularly when considering ethnic groups or sub-groups, the common element, or g-factor tends to dominate.

This would seem to imply that work with the highly selected group of able pupils should certainly take particular account of specific abilities, and that it is entirely appropriate that a place should be found in definitions of giftedness for the pupils who have high ability in relation to one particular kind of ability even if they do not, at the same time, meet some (arbitrary) IQ criterion as well.

There is an important implication for identification if we accept the view that we should base our strategy on a specific rather than general theory of ability. This concerns the number of pupils who will fall within the scope of the two different kinds of definition of 'giftedness' which reflect the two points of view. Ogilvie (1973) explored this in an interesting theoretical way. He supposed, for example, that we consider eight school subjects and that abilities in these subjects are correlated with a correlation coefficient of 0.5. If lists are then drawn up of the top 5.5% in each subject, 55 pupils in every thousand will appear on each list. Ogilvie showed that under these circumstances only 1 pupil out of the 1000 would appear on all eight lists. Clearly if we defined giftedness in terms of high ability across the full range of subjects, many pupils with high ability in one subject would be denied membership of our gifted group. If, on the other hand, we counted any pupil who appeared on any of the top 5.5% lists as gifted, then the overlap between the lists would imply that

235 separate individuals out of the total number of 1000 individuals will be counted as gifted. There is support for this theoretical perspective from the following tables which have been compiled from results which we have collected in our own work.

Table 1.1 shows the percentages of children selected by teachers as being in the top 10% in terms of O-level potential in one particular subject, who were also selected as being in the top 10% in another subject. Thus, for example, 56% of the pupils categorised (on this criterion) as more able in English were also categorised as being more able in French.

TABLE 1.1

THE PERCENTAGE OF PUPILS NOMINATED BY TEACHERS AS BEING IN THE TOP 10% IN ONE SUBJECT WHO WERE ALSO NOMINATED IN ANOTHER SUBJECT

Nominated In	Percentage also nominated in		
	FRENCH	PHYSICS	MATHS
ENGLISH	56%	52%	49%
	ENGLISH	PHYSICS	MATHS
FRENCH	61%	46%	56%
	ENGLISH	FRENCH	MATHS
PHYSICS	51%	54%	50%
	ENGLISH	FRENCH	PHYSICS
MATHS	48%	54%	51%

These figures are based on work done in eleven schools with approximately 2300 third form pupils. The method of selecting pupils which gave rise to these figures will be described in greater detail later, but with Ogilvie's result in mind it is interesting to note that there was indeed a substantial lack of agreement between the lists of pupils recognised as 'more able' in one subject, and the corresponding list for another subject.

TABLE 1.2

THE PERCENTAGE OF PUPILS NOMINATED BY TESTS AS BEING IN THE TOP 10% IN ONE SUBJECT WHO WERE ALSO NOMINATED IN ANOTHER SUBJECT

Nominated In	Percentage also nominated in		
	FRENCH	PHYSICS	MATHS
ENGLISH	63%	45%	55%
	ENGLISH	PHYSICS	MATHS
FRENCH	63%	40%	56%
	ENGLISH	FRENCH	MATHS
PHYSICS	45%	44%	67%
	ENGLISH	FRENCH	PHYSICS
MATHS	53%	54%	66%

Table 1.2 indicates that identification based on test measures of pupils' potential for O-level performance in these subjects gave rise to very similar results in terms of the amount of overlap between top 10% lists in the four subjects. Again these results would seem to support Ogilvie's theoretical model. Table 1.3 completes the picture.

TABLE 1.3

THE PERCENTAGE OF PUPILS NOMINATED AS IN THE TOP 10% IN ONLY 1, 2, 3 OR 4 SUBJECTS

No. of Subjects	1	2	3	4
Teacher Nomination	45%	23%	20%	12%
Test Nomination	40%	30%	15%	15%

It shows the percentage of the total number of pupils nominated in any of the subjects who were nominated in only one, in two and only two, in three and only three, or in all four subjects. Again teachers' nominations and test-based nominations are shown. There is again considerable agreement between teachers and tests and the pattern predicted by Ogilvie is apparent. If assessment were to be restricted to a smaller ability band or if the numbers of subjects were to be increased, it is interesting to conjecture that very small numbers would appear in the cells indicating high ability on all or most subjects.

Ogilvie's theoretical figures, supported by our experimental results indicate strongly that the decision to base an identification strategy on a general or specific view of ability will have significant consequences for the total number of children who will be drawn into any programme of provision that is subsequently mounted. If one also bears in mind that teachers (and HMI) wished that provision be made for a fairly large proportion of children in each subject, one can see that the total number of pupils drawn into such a programme of provision in a school may be very much greater than the tiny minority traditionally regarded as 'gifted'.

To summarise the position so far: it would seem that there is considerable support, both in the developing definitions of giftedness and in the psychological literature on the structure of ability, for the appropriateness of the subject specific aspect of the work which we have been urged to do. There also seems to be considerable evidence that significant numbers of children will indeed appear as more able in only one or two subjects.

We now turn to the literature which sheds some light on the particular form of identification - teacher based identification - on which we were asked to concentrate. We look first at literature which discusses the reasons why teachers should be a main component of the identification process.

The extent to which British projects rely upon this approach has already been noted. A working party of the City of Birmingham Education Committee (City of Birmingham, 1979) presented a very well argued case for this reliance. They stressed the need to see identification as a continuous process of re-assessment of childrens' needs and emphasised that, because of this, the teacher would always be a major component in the identification system. One can find direct confirmation for this view in the work of Renzulli, Reis and Smith (1981) in America. They suggested a 'Revolving Door Model' for identification which stressed the idea that giftedness was not something to be assessed on a once-and-for-all basis but rather that children should move into and out of any special programmes for the gifted as their own needs in relation to their normal educational curriculum change. Renzulli and colleagues argued that the ordinary class teacher must be central to a system based upon this model. One can certainly see that a battery of tests large enough to provide an adequate map of the full range of abilities is unlikely to be short enough to permit the frequent use that would be necessary if it was to form the basic identification instrument in a 'Revolving Door' system.

To these central points can be added the argument that, if provision for the more able is, at least in part, to be in the hands of the ordinary classroom teacher then it is essential that that teacher has confidence in the identification system that brings children into the programme of provision that she/he is organising. We would suggest that if a large battery of tests is used, the interpretation of results is likely to be complex. (We discuss this in greater detail in a later chapter.) A list of children generated in some complex way from such a battery is unlikely to be used with confidence by the teacher. For this reason alone it would seem to be sensible that the teacher's own judgement should be a major part of the identification scheme. Martinson (1975) makes a similar point and also goes on to suggest that there is a useful general 'spin off' as 'participation in study and observation of the characteristics of the gifted is valuable in-service training'.

These would seem to be some very positive reasons for suggesting that the widespread dependence on teacher based identification is essentially sound. If teachers have difficulty in identifying appropriate pupils accurately, we should therefore seek to improve their ability rather than to turn immediately to an alternative method of identification.

It would seem, at this point, that we should attempt to look in far more detail at what the literature has to say about the extent of teachers success in identifying pupils who are able in specific subjects.

An important paper which discusses teacher based identification of gifted pupils is that of Pegnato and Birch (1959). They chose a definition of giftedness which was based on the scores achieved by pupils on the Stanford-Binet IQ test, administered individually to pupils by an educational psychologist: pupils scoring 136 or more on this test were considered as 'gifted'. In principle, working with such a definition, identification could be carried out with a high level of precision by individually testing all children in the schools. However they pointed out that 'few schools in this nation (USA) have access to sufficient psychological services to provide for individual examination for all children' and therefore suggested that, in practice, identification should be carried out in two stages: first the whole school population should be screened by means of group tests of IQ and of achievement. On this evidence appropriate pupils should be nominated for membership of a 'talent pool'. Secondly these 'talent pool' pupils should be tested individually, by a psychologist, to determine which of them were indeed gifted. A similar procedure was recommended by Gallagher (1966) after his comprehensive survey of the reseach on identification which was then available.

In view of the definition of giftedness used by Pegnato and Birch, it is natural that they should have recommended individual testing by psychologist as the second stage of their identification process. Their choice of group tests of IQ and achievement for the screening stage was based on two considerations: first that the screen must be effective (i.e. that it must place, in the talent pool, all of those pupils who would have qualified for the label 'gifted' if everyone had been given the individual IQ test); and secondly that it should, as far as possible, be efficient (i.e. that it should minimize the number of 'non gifted' who would also be placed in the talent pool). Pegnato and Birch considered the efficiency and effectiveness of several screening procedures. They showed that, used together, group tests of IQ and achievement were 97% efficient and concluded that they provided the best means of screening. They also showed that, in the context of their study, teachers' nominations were neither effective nor efficient enough for them to be able 'to place much reliance on them for screening'. (They found teachers to be 45% effective and 27% efficient). Many other studies, using similar IQ-based definitions and following a similar methodology, have come to the same conclusion with regard to teacher nomination. Useful summaries of these are given by Jacobs (1971), and by Martinson (1975).

Some important work on approaches to provision for the gifted has made use of IQ-based definitions of giftedness (Bridges 1969, Tempest 1974, Bridges 1975) and, of course, for such work, the finding of the researchers quoted above will be both pertinent and useful. However, we have argued that the high IQ group should be thought of as only one small sub-set within the broader definitions of giftedness for which we have tried to make a strong case. This of course changes the perspective in which we should view the work of Pegnato and Birch and the other authors of similar studies. No longer should their findings and recommendations be seen as prescriptions for the identification of 'the gifted', but simply as recommendations for the identification of the one sub-set of the gifted which is defined in terms of high IQ. Of course, to ignore their findings in relation to this sub-set would be to act unwisely. However it would be equally unwise to generalise their findings, uncritically, to other sub-sets such as that containing children with high subject specific ability.

Therefore, while we must keep such studies in mind, we must also recognise that they do not provide direct answers to the questions which we wish to tackle. Indeed the literature seems to provide no detailed assessment of the effectiveness of teachers in identifying pupils with high subject specific ability. There would seem, therefore, to be a very real need for the kind of study which is described in the rest of this report.

One other part of our study needs to be located in relation to the relevant literature. We made a case earlier for paying particular attention to the value of checklists in guiding teachers judgement. This case is clearly strengthened by the suggestions set out above: first that teachers should be involved, and secondly that they may not (if Pegnato and Birch's work does prove to be generalisable) be very effective nor very efficient. We therefore attempted to trace how the checklists that have been offered to teachers have been constructed, and to seek evidence of their effectiveness in improving teacher based identification.

The usual approach in constructing a checklist has been to look to research to provide evidence about the characteristics of pupils so that teachers could be encouraged to look for appropriate clues. This very reasonable approach has led to the production of many essentially similar checklists. Shields (1973) has listed several of these. A further example is the list devised by Laycock (1959) and quoted as recently as 1980 in a D.E.S. pamphlet for the guidance of teachers. In the context of our subject specific approach to identification there is a serious pitfall in this procedure. It is that many of the insights to be gleaned from earlier research are insights into the characteristics of the high IQ children who were the subjects of that research. Thus the checklists generated on the basis of such insights will, in principle, be of dubious worth for the purposes of our study.

Freeman (1981) provides warning of another kind of problem which she argues is inherent in many of the general checklists that have emerged in recent years. She writes:

> Some educationalists prefer to avoid IQ tests by, for example, drawing up lists of characteristics. In time the lists grow to accommodate the inevitable exceptions until they become so all embracing that they would describe almost any child. Unfortunately the lists have not been tested for their validity or predictive value.

A further problem arises from the nature of the items which make up the lists. Much of the research that gave rise to the items was not classroom based. It is therefore possible that, valid though these items might be as correlates of high general abiltiy, they may not actually be observable in the classroom situation. Clearly what can be observed of a child in a study which devotes all of its efforts to the observation of an individual may not be suitable for direct use in the classroom where direct observation of one idividual cannot be carried out so intensively. (Indeed the clues that may be observable in class and that might point to a child as having some important trait might be more appropriate checklist items than a broad statement of the trait itself.) Certainly the ease with which a teacher can complete a checklist will be important because, in a busy classroom, this will be only one aspect of what the teacher has to do. When teachers are confronted with a checklist on which the items are too obscure for easy observation they may well lose confidence in the instrument and revert to dependence on their mark book and their general impressions of the pupils so that the checklist, though it may actually be completed, may have only a very small role in forming the teacher's final opinion of a child.

Thus there would seem to be several problems in the use of general checklists for the guidance of teachers, and especially for the guidance of teachers in making subject specific identifications. It is therefore encouraging to note that, as well as listing several general checklists, the booklet on identification which has recently been produced by Schools Council Programme 4 also contains a strategy which teachers might use in compiling their own list for their own particular purpose (such as subject specific identification). The method, which is called 'Up the Pole' (Clarke, 1981) is, however, merely a systematic way by which a group of teachers can agree on what items they will include on their lists. However, the items that they consider for inclusion are drawn simply from their own experience so the lists will not necessarily contain items which are valid indicators of ability in a subject. The items will be relevant to certain activities within the subject but will be based only on subjective judgements which will themselves be limited by individual experiences of the teachers involved. These experiences may themselves reflect the teachers' limited aims in their subject teaching.

Such checklists containing the sorts of items which might emerge through the 'Up the Pole' method have been used in Science in the selection of pupils for special provision. The 'Cleveland' Checklist that is reported in the Schools Council booklet contains a variety of such items but it would require considerable research in order to ascertain which of them were particlarly relevant to Science and which were equally relevant to other subjects. For example one of the items 'Has a wide range of interests and/or several hobbies, reads and talks accordngly' is the kind of ability associated with the general checklists. Another, 'Possesses an unusual imagination', could be important for Science and could be demonstrated to be so by research, but the item may not necessarily be valid. It could be argued that an unusual imagination could hinder a child in accepting some basic scientific principles and that, in consequence, though he may be a creator of new ideas, he might not be able to acquire a sound basis of knowledge from which the new ideas could come. These particular items are included on the basis of a certain kind of reasoning and the point is that the reasoning well be unsound.

Straker (1981) in her companion booklet dealing specifically with Able Mathematics pupils approaches the identification strategy with caution, quoting a wide range of items from behavioural criteria, applicable to able children in general, to specific items relating to particular aptitudes within mathematics. The teacher is left with the task of extracting what seems appropriate. This reflects the current position of uncertainty as to what is necessary and sufficient for a checklist relevant to a subject such as Mathematics. No doubt many of the items would be valid from these lists and some will have been included as a result of careful research, but this is only conjecture.

These examples illustrate that whereas the general checklists contain quite well researched items but reflect an IQ bias, the potentially more interesting subject specific checklists tend to contain items of unproven validity. This unproven validity of individual items may not be a catastrophic problem if the checklist as a whole could be shown to be effective in guiding teachers to make valid judgements about pupils. However very few studies have been conducted into the overall effectiveness of any form of checklist. Renzulli, Hartman and Callahan (1971) provide a detailed report on attempts to validate the 'Scale for Rating Behavioural Characteristics of Superior Students'. This scale (Renzulli and Hartman, 1971) is very well constructed on the basis of research and contains four sub-scales (Learning Characteristics, Motivational Characteristics, Creativity Characteristics and Leadership Characteristics). They showed that there were high correlations between ratings on these sub-scales and other objective measures of the relevant characteristic. Solomon (1979) also makes passing reference to the effectiveness of a checklist in improving teachers' ability to identify 'intellectually gifted fourth year junior pupils'. These studies are intriguing in that they provide some general support for the idea that checklists may be worthwhile. However, we have been able to find no detailed information on the effectiveness of any checklist which is focussed on a specific subject ability.

Our survey of this part of the literature only serves, therefore, to increase our conviction that a study of subject specific checklists is very much needed at the moment.

1.4 Summary

We have attempted above to show how the broad direction of our project was defined by discussion with teachers and visits to other projects which are working with more able pupils. We have also tried to establish, by reference to the literature that this broad direction was a sensible one. What has also emerged from the survey of the literature is that the answers to the detailed questions which we set out at the end of Section 1.2 are not to be found in currently published work. We feel therefore that a clear justification for the present project has been established and will now turn to a description of the methodology which we have used in taking this work forward.

CHAPTER TWO

Research Design

2.1 Introduction

In Chapter One we listed five questions which emerged from our discussions with practitioners and from our review of the literature. In this chapter we will discuss the broad outlines of the research design which we have adopted in an attempt to provide some answers to these questions. We will also discuss, in more detail, some of the practical aspects of the work which was undertaken as a preliminary to the main study.

2.2 Outline of the research design

Before any work could be undertaken in schools it was necessary to focus, much more sharply, the questions which we have decided to study. Key points to be considered in achieving this sharper focus were:

- which age group of pupils should be studied?
- which subjects should be studied?
- what proportion of children should be taken as 'more able' for the purposes of the study?
- what, exactly, should be taken as the criterion used for labelling pupils as 'more able' in a given subject; and how should the effectiveness of teachers in identifying such pupils be assessed?
- how can the nature of the process by which teachers identify 'more able' pupils be explored so that the kinds of things which lead them to make errors can be revealed?
- how should the effectiveness of checklists as guides for teachers be assessed?
- what types of schools should be studied and compared?

It is through answers to this set of questions that we hope to bring the reader to an understanding of the overall research design. We begin with the question of the age group of pupils to be studied.

If the case for continuous assessment of pupils, made in Chapter One, is accepted, an argument can clearly be advanced for a study of any, or even of all, age groups of children. We decided that a detailed study of one age group within the comprehensive secondary school would be more useful than a necessarily more superficial study of all age groups. We then chose to focus our attention on the third year (13+ pupils). We rehearse below several reasons for this choice - though as we have already suggested, we would wish there to be no implication that this should be the only year in which teachers in comprehensive schools should attempt to identify the more able.

First, it is common practice for pupils to make their final selections of optional subjects during their third form. This is, therefore, a critical year for the assessment of pupils' abilities by subject teachers, so that proper guidance can be given. The need for accurate assessment is strongly felt in all schools, and particularly so in the upper schools where 13-14 year olds are embarking on their first year at the school after moving up from middle schools. Thus in asking teachers to identify more able pupils in the third year we were asking them only to engage in an extension of a process which is already accepted as a sensible and necessary part of the normal routine of the school.

Secondly the third form is where some subjects, such as the three sciences, are studied as separate subjects for the first time. If work were to be done in earlier years when the sciences are often taught in some form of combined or integrated way, it would be difficult to draw a representative of this important subject group into the study. A similar argument can be applied in another important subject area: namely English. There is considerable variation between schools in the extent to which the teaching of English is integrated with the teaching of other subjects in years one and two. If our work had been in these years, this variation would make it difficult to interpret comparisons made between schools. The situation, in these and in other areas of the curriculum, is much more clear cut in the third year.

Thirdly, by year three, the strengths and weaknesses of pupils are becoming more apparent as the levels of challenge in each subject increase. This level of challenge is related, to some extent, to the fact that at around the age of 13 some 20% of children will be operating firmly within the Piagetian stage of Formal Operational Thought (Sutton and Haysom, 1974). This 20% would certainly include all pupils who might reasonably be regarded as more able. 13+ is therefore an age when it should be possible for teachers to observe clues relating to the potential of more able pupils to handle formal as well as more concrete aspects of individual subjects. The force of this particular aspect of the argument is enhanced by the fact that our study will be concerned with a fairly large target group. Studies concentrating on a tiny proportion of the top end of the ability range could reasonably argue that children in their target group will be capable of Formal Operational Thought, and will therefore display relevant clues, at a much earlier age. On essentially this same point, one could suggest that the third form is an especially appropriate time for enrichment of the educational diet for these children: a time when the levels of challenge that could be met by children of different abilities are drawing noticeably apart. If this is accepted, 13 is clearly an age when it is particularly appropriate to try to identify the more able pupil.

There were also some pragmatic reasons for the choice of year group to be no later than the third form in our research. In view of the final two years of preparation for public examinations it would not have been acceptable to those who cooperated with us to impose a large research effort on pupils in their 4th or 5th years.

As we have suggested, a case could be argued for a study of any year group in the secondary school but we feel that the points above make the choice of the third year particularly reasonable. Certainly, by concentrating the effort on the third form we hope that not only will the depth of reseach be possible, but also that results can be generalised to some extent to other year groups.

We turn now to the question of the number and the nature of subjects to be studied. Again it is clear that if we were to spread the research effort too widely we would limit the depth that could be achieved. However we would argue that the choice of subjects should surely be wide enough to enable insights to be gained into more than one key area of the curriculum. We decided that four subjects would provide a reasonable compromise in the face of these conflicting demands. Two subjects, mathematics and English, were chosen because there is common agreement about their importance, both in their own right and as service subjects for other parts of the curriculum. It would have been difficult to make a strong case for any other single curriculum area on the basis only of its intrinsic importance. Therefore, to get a range of insights, we chose areas which were certainly important in themselves but which also provided considerable variety in the circumstances in which the teaching was usually carried out. Thus we chose a science (clearly not an unimportant area in its own right, but one which was also chosen because of its emphasis on practical work and on long blocks of teaching time, and because of the relatively infrequent contact between teacher and class) and we chose a modern language (again important, and different from the others - especially mathematics and science - because of its emphasis on oral work). Physics was taken as the example of the sciences, and French of the languages. We would hope that many of our major results could be generalised to related subjects, but would not pretend the conclusions are necessarily valid in all areas of the curriculum. They may well not be valid, for example, in practical or aesthetic subjects, or in relation to the physical activities that are undertaken at school.

The next issue to be considered is that of the size of the target group. In Chapter One we argued the case for adopting a fairly large proportion of pupils of a particular age as our target group in each subject as this would reflect the concern, both of HMI and of the teachers whom we interviewed, that far more than the top 2% who are traditionally regarded as 'gifted' have need of some (perhaps fairly limited) additional, or alternative provision. In the absence, at present, of any strong theoretical reason for supposing that a particular proportion of pupils at the top of the ability scale are qualitatively different from other pupils in terms of subject specific ability, the decision as to exactly how big this proportion should be was necessarily rather arbitrary. We actually decided to work on the 'top 10%' in each subject. A group as wide as 10% will of course include the top 5% and other smaller sub-groups, and the research was structured so that relevant work could be done on these narrower target groups that are certainly of interest in some situations (e.g. when selecting a very few pupils to take part in a summer school). In the schools in which we worked, the top 10% was of the order of 15-20 pupils, who would therefore comprise the top half of the top teaching set in a year group of a school that divided its pupils into ability groups at third form level. An important decision was that to define the group relative to their peers in their own school. This decision was taken as the needs of particular children will clearly be dependent upon the context in which they are educated. Although the 'top 10%' in one school may differ in absolute ability from the 'top 10%' in another, it is likely that both groups, being well above average in their own environment, may have need of some special provision. This point was discussed at some length, and a similar conclusion was reached, in the report of the City of Birmingham study group (1979), to which we referred in Chapter One.

The main reason for the decision to use school norms rather than national norms was explained above. However it is important to add a further point. Asking a teacher to identify the top 10% on the basis of national norms is to ask the almost impossible, as the full information necessary for the teacher to undertake such a task successfully, simply will not be available unless the teacher has detailed knowledge of the abilities of children from all areas of the country. To devote time to demonstrating that teachers are relatively ineffective at carrying out an almost impossible task seems to be unwise when the same resources can be used to explore the more interesting, more realistic and, we would suggest, actually more important issue of identification on school norms.

We now turn to the crucial issues of what criterion is to be used as the basis for labelling a pupil as a member of the top 10% group in one of the four subjects, and of what method can be used to assess the effectiveness of teachers in identifying the pupils who match this criterion. If some judgement is to be made about the effectiveness of teacher-based identification it is neccessary that one should be able to define the target group which the teacher is trying to identify in terms of some clear criterion. It is also necessary that one should ensure that the teacher is completely clear about this definition. Once this has been done and the identification has been made, one must compare the teacher's judgement either with the criterion itself, or with some other measure of it. Let us consider an example. We might ask a music teacher to identify pupils who have already reached Grade Five on the oboe. We would need to be sure that the teacher knew exactly what he was being asked to do. We could then measure his effectiveness by comparing his list of names with the names of pupils who had a Grade Five certificate.

As another, and perhaps more interesting example, we may wish to assess the effectiveness of the teacher in identifying those 13 year old pupils who have the potential to achieve Grade Five oboe by the age of 16. In this case, judging the effectiveness of the teacher is more difficult. We would need to seek out some alternative means of assessing potential for playing the oboe. We might use the subjective judgement of another, more expert judge (or panel of judges) or we might look for some objective test that has already been shown to be a good predictor for future success in the Grade

Five oboe exam. With both of these alternative measures we would need to take some account of the possibility that our alternative measure of potential may not always be accurate and that some of the mismatches between the teacher's list and the other list may not actually be indicative of error on the teacher's part. Taking account of the error in the alternative measure is an important part of the analysis of the results of such a study. Another absolutely essential thing is that there is consistency between the nature of the task we ask of the teacher and the nature of the characteristic that is being measured by the alternative method against which we are judging the teacher; thus present Grade Five results are useful in the first example but probably of very little use in the second.

As long as the consistency, to which we have drawn attention above is present, valid if perhaps very limited results can emerge. However, in an attempt to extend the scope of the findings the researcher - or an individual reader - may seek to imply that (to continue with the example) Grade Five oboe is an excellent measure of overall musical ability and that the study can be used to give insights into the effectiveness of teachers in identifying pupils who are of high ability in terms of this much broader characteristic. It is important to recognise that we can accept or reject the validity of these broader claims for the study without questioning its value as an investigation of the original, and narrower question to do with Grade Five oboe potential.

In our study it was decided that the criterion measure in each of the four subjects should be O-level performance in that subject. Thus we asked teachers to identify that 10% of third year pupils who, relative to other pupils in their year group in their school, had the greatest potential for future achievement at O-level in that subject. We then sought out a battery of tests which, with known error, could provide an alternative prediction of which 13 year old pupils would be in this top 10% group in terms of their future O-level performance in that subject, in that school. After taking due account of the error associated with the test-based prediction, we assessed the effectiveness of teacher-based judgement by comparing the teachers' lists with those generated from the test scores. We suggest that this procedure meets the essential criterion of consistency that we drew attention to above, and that it can be supported as reasonable by two further arguments:

i) that the top 10% target group lay well within the ability band for which the O-level exam is designed;

ii) that the O-level criterion is in line with the basic aims of the third year teaching so that we could reasonably expect teachers to have appropriate information available to them on which to base their judgements.

We would suggest, therefore, that the chapters which follow provide valid insights into the question of how effectively teachers can identify pupils who have high potential for future O-level performance in a given subject.

Some readers may wish to speculate on whether or not the results also show the extent to which teachers can identify, for example, 'able mathematicians' in a much more general sense. We offer some comment on this below, though we would stress that whether or not the reader wishes to make such generalisations, the findings will still be valid in the more limited O-level context.

Some support for the view that high ability at O-level can be used as an indicator of high ability in a broader sense can be found in work which attempts to measure the predictive validity of O-level itself. For example evidence suggests that O-level performance in a subject is highly correlated with later performance on more demanding criteria in the same subject. Murphy (1981) has, for example, found the following correlations between O-level and A-level mark:

TABLE 2.1

Subject	No. of pupils	Correlation[1]
ENGLISH (Language)	3462	0.63
FRENCH	1688	0.90
PHYSICS[2]	3407	0.81
MATHS	3110	0.81

This line of argument can be continued, though with admittedly attenuated force, by noting that A-level has been found to be an important factor in predicting degree success, particularly for Science students (Entwistle and Wilson 1977).

The reservations associated with generalising the O-level criterion are most strongly held in English Language. Willmott and Nuttall (1975) reported that the English Language examination gave the lowest reliability of all subjects (English Language 0.70, French 0.85, physics 0.89, mathematics 0.94). An earlier DES (1964) document discussed the same problem and this was amplified by Stevens (1970), though it was also reported by the DES that no other factor had affected syllabus design in English more than the O-level examination nor was there any better standard of assessment available.

A final point is worth making. One might argue that what is often seen in the work of pupils taking O-level is very limited compared with the challenges which could be met by the more able pupil. This criticism may indeed be valid in that it may be possible to achieve a pass at O-level by exercising a limited range of fairly low level aptitudes (such as direct factual recall). Yet it is the narrowness of the teaching that is being criticised here, rather than the O-level examination itself. Many would suggest that the O-level syllabus is designed to allow breadth and depth of challenge to those who are able to respond, and that to achieve a high grade, (and more particularly to achieve a high mark within the top grade) a pupil will need to exercise much more than the limited abilities required for a mere pass. The development of our criterion measure has involved the use of the full range of O-level scores and can, therefore, be considered as measuring potential for these higher level challenges.

We have reasoned why O-level potential was taken as our background criterion, so it remains to explain how prediction for O-level success was achieved. It could be argued that a series of tests specific to the individual subjects should be administered. The problem is, however, that tests of this kind which have known predictive validity for O-level do not exist and it would take a major project to develop them.

An alternative approach was available, however, in that we had, for the purposes of previous work (Postlethwaite and Denton, 1978), and after consultation with psychologists at NFER, tested a number of children with a group of tests made up from the Differential Aptitude Test Battery (Bennett et al, 1974) together with measures of attitude[3] and creativity[4]. We therefore decided to explore the use of this group of tests in detail and to establish appropriate combinations of test scores for each subject which could be used as predictors of O-level performance in those subjects. We were able to achieve this because data from the same group of tests was available on two groups of 13 year olds in two schools. Roughly a hundred of these pupils who were in their 3rd year in 1974 and 1975 had gone on to take O-level examinations in each of the subjects, but no attempt had been made to identify individuals as more able. It was therefore possible, without any distortion arising from the well documented relationship between expectation and performance[5] to establish by means of multiple regression analysis, relationships between pupils' test scores at 13 and their subsequent O-level performance in each of the chosen subjects. The resulting prediction equations, when applied to the data from subsequent groups of 13 year olds in this present study, were used to generate pupils' predicted O-level scores. It is these predicted O-level scores which we have interpreted as objective measures of potential. Appendix 1 contains details of the derivation of the equations.

An important consequence of this particlar test based criterion was that not only did we have subject specific tests of ability, but also estimates of the errors involved in using the test results as estimates of future O-level performance. This is because the multiple regression method yields the best combination of tests for the prediction required and also the average error of that prediction, based on the scores of the sample from which the equations are derived. In any test there is error. By having an estimate of the error we can make allowance for it when pursuing investigations to which the test scores are related.

As will be discussed later it has been possible, in part of our study of mathematics, to use as an alternative background criterion the present aptitude levels which are displayed by the pupils on the spectrum of abilities that are relevant to work in mathematics. An implicit assumption is made in measuring these present aptitudes: namely that if a child has high scores on this spectrum of abilities, then he/she has potential which could be developed.

We are then, for the greater part of the work, using the O level criterion albeit with some reservation in connection with its generalisability, but in so doing, we are seeking answers to an important specific problem the accuracy of teachers in identifying pupils for the highest challenges that are set by their present course of study. The relevance of that to an identification concerned with a broader concept of high ability in the subject can only be a matter for speculation, though as we have suggested there is a certain amount of supportive evidence available.

The next issue is that of how to gain insights into the process which leads teachers to make judgements about whether or not a pupil was in the 'top 10%' group. This was tackled in three ways. First we made comparisons, on the basis of the individual background test scales (rather then the combinations of scales which gave us our predicted O-level measure), between pupils who were identified by their teachers and pupils who had similar potential for the subject but who were overlooked by their teachers. We looked, for example at the numbers of boys and girls who were identified and at the numbers who were overlooked to see if there was any evidence that the sex of the pupil played an important part in forming the teacher's overall judgement. This work is described in detail in Chapter Four. The second, and more direct method of enquiring into the process which led teachers to their overall judgement involved interviews with them in which we asked them to talk about the characteristics of particular pupils. By suggesting the names of pupils who were identified and of similar pupils who were not, we gained some insights into the kinds of pupil characteristics which teachers noticed, and were able to form some tentative views about which of these characteristics tended to be associated with underestimation, and which with overestimation of overall potential in the subject. The third method was a more rigorous development of the interview technique and drew upon the principles and methodology of personal construct psychology (Kelly, 1955). Later chapters describe this work in detail. We devoted considerable effort to this aspect of the study as we were clear that only through an understanding of the process could we hope to offer any advice on the question of how teachers could become more effective in identifying more able pupils. Indeed it is the lack of such insights which severely limits the usefulness of many of the other studies of teacher-based identification that we referred to in Chapter One.

So that we could assess the effectiveness of checklists as a form of guidance for teachers we decided to adopt the following strategy. In the Autumn Term all schools were asked to identify more able pupils without any guidance from checklists. Schools were then divided at random into experimental and control groups. In the Spring Term all schools were asked to make a second identification. This time the experimental schools used the subject specific checklists, whereas the control schools again worked without any such guidance. Comparisons between the two groups of schools, We would remind the reader that there is quite strong evidence that O-level performance, particularly in physics and mathematics, is a good predictor of performance at much higher levels in the subject. The reader may therefore wish to generalise our results from the specific issue of the comparison of test based and teacher based identification of high O-level potential to the broader issue of test based and teacher based identification of potential in a much more general sense.

The identification was intended to be from the whole of the 3rd form, not from any particular subgroup such as the top set in the subject. The results of the simple comparisons between the teacher-based and test-based groups are set out in Table 3.1 below.

TABLE 3.1

PERCENTAGE OF TEST IDENTIFIED PUPILS ALSO IDENTIFIED BY TEACHERS
(Number in group is given in brackets.)

SCHOOL	ENGLISH	FRENCH	PHYSICS	MATHS
00	57 (14)	69 (13)	44 (16)	64 (14)
01	67 (18)	52 (23)	43 (21)	59 (17)
02	53 (19)	58 (19)	52 (21)	65 (20)
03	75 (15)	50 (12)	50 (14)	36 (14)
04	67 (18)	25 (16)	20 (20)	53 (19)
05	56 (16)	62 (16)	53 (15)	78 (14)
06	70 (20)	50 (20)	48 (23)	50 (20)
07	50 (22)	38 (21)	37 (16)	67 (24)
08	75 (12)	50 (12)	17 (12)	58 (12)
09	71 (24)	62 (24)	54 (24)	67 (24)
10	67 (12)	73 (11)	75 (12)	50 (12)
11	55 (11)	45 (11)	20 (10)	64 (11)
12	33 (12)	25 (12)	58 (12)	64 (12)
13	51 (35)	47 (34)	53 (34)	59 (35)

NOTE: At 3rd year level, school 13 exists in three fairly distinctunits which are also analyzed separately as schools 10, 11 and 12 in the above table.

When similar results were calculated for the sample as a whole the figures shown in Table 3.2 were obtained.

TABLE 2.2

		13 - 18 Age Group		11 - 18 Age Group	
		COUNTY	SAMPLE	COUNTY	SAMPLE
1000 or less	BOYS	1	1	1	0
	GIRLS	1	1	0	0
	MIXED	5	1	15	2
1000 - 1500	BOYS	0	0	0	0
	GIRLS	0	0	1	0
	MIXED	1	0	15	5
1500 + pupils	BOYS	0	0	0	0
	GIRLS	0	0	0	0
	MIXED	0	0	2	1

(ONE SCHOOL IS 11-16 WITH 1000 OR LESS PUPILS)

TABLE 2.3

	COUNTY	SAMPLE
MIXED	38	9
SINGLE SEX	4	2

TABLE 2.4

	COUNTY	SAMPLE
13-18	8	3
11-18	34	8

(ONE SCHOOL IS 11-16)

TABLE 2.5

	COUNTY	SAMPLE
1000-	23	5
1000+	19	6

It turned out that the school with more than 1500 pupils was made up of three lower school (11-15 yrs.) halls and an upper school (16-18 yrs.). In many ways these halls functioned as separate schools. We took account of this in our analysis when appropriate, so that in the report we will sometimes be showing figures for 13 schools rather than 11. Further details of the schools and resultant sample of pupils is contained in Appendix 3. These will enable the reader to form some view about where the pupils stand in relation to various national norms. While we would not wish to claim that the schools are representative of all kinds of schools in England and Wales (e.g. there is no sense in which schools serving very deprived inner city areas are represented) we would suggest that the sample is less strongly biased towards the 'advantaged' end of the social and ability spectra than might be supposed from superficial thought about the nature of the county of Oxfordshire.

The approximate numbers of pupils tested for the study were 2400 in the first year and 1500 in the second year.

It is perhaps helpful to point out that the sample of schools enabled us to make several interesting comparisons: e.g. between schools having 11+ entry and those having 13+ entry; and between schools of very different sizes. Some comments can also be made about differences between mixed and single sex schools.

When a project reaches its final stages it is interesting to look back and discover how closely the investigation strategy followed the initial design. Inevitably one accumulates information as a project proceeds which leads to a change of emphasis for the later stages, so that the research design, in some ways evolves with the project. To some extent that has been true of this project. The first year's work became a large computer orientated statistical exercise concerned with accuracy of teacher identification compared with the test based criterion related to O-level and the effectiveness of checklists in guiding teachers. We were careful to take account of test error as far as possible and this led us to reflect quite deeply on what we really understood by accuracy of teacher judgement. This reflection emphasised the importance at looking more closely at the process of identification generally, and at the processes involved in using checklists. We were thus moved to look more closely at the classroom aspects of the identification process than we originally intended, during the second year. This new emphasis for Year Two was possible because of the completeness of the first year's work, which relieved us of the task of repeating certain investigations to confirm what we had found. It reflected the fact that we had moved away from our original standpoint of an objective comparison of test and teacher identification that set the test up as the ultimate criterion. We will now summarise the research design that evolved in this way.

Preliminary work

a) Decide upon major issues for the research by visits and reading.
b) Establish prediction equations.
c) Select sample of schools.

Year 1

a) Early in Autumn term. Administer tests to all third form pupils in the representative sample of schools. These tests form the basis of the background criterion because of their relationship with O-level performance in each subject.

b) Autumn Term. Ask teachers in all 11 schools and in each of the 4 subjects asked to nominate the 'more able' (top 10%) group.

c) Intensive statistical work to assess extent of agreement between tests and teachers and to look for characteristics from the background tests that give clues to the criteria that the teachers tend to use. Analysis to be according to school and type of school.

d) Schools randomly assigned as control or experimental.

e) Early Spring Term. The strategy of using subject specific checklists to be discussed in experimental schools.

f) Late Spring Term. In all schools a second identification to be made within the same 3rd forms. In the control schools this is to be unaided, in the experimental schools it is to be assisted by the checklist records.

g) A second statistical analysis is to be made comparing round 1 identification with round 2 in control and checklist schools and control and experimental for round 2 to assess the effect of checklists on the identification process.

h) Summer Term. Interviews with teachers to look for clues to their identification strategy to complement the statistical evidence, in particular to look for reasons why some pupils are categorised differently by test and teacher.

Year 2

a) Early Autumn Term. Repeat of testing programme for new 3rd forms. Continuation of interviews.

b) In each subject a sample of schools to be chosen for more intensive checklist work and another sample for the personal construct work[6].

c) Late Autumn Term. 'Personal Construct schools' make identifications of 'more able' group.

d) Early Spring Term. Checklist schools begin the identification process, the method of use being a development from Year 1. Classroom observation begins[7] in order to accumulate deeper insights into the identification process.

e) Spring Term. Personal Construct Analysis developed and used to form basis of interviews with teachers co-operating in this work, the aim being to discover what characteristics influence the teacher in his/her judgements.

f) Late Spring Term. Identifications made by teachers who have been using checklists. These to be compared with test based assessment and also comparisons to be made with the previous year's results so that final recommendations can be made.

Notes for Chapter Two

1) These correlations have been corrected, by Murphy, to take account of the limited range in the O-level scores of students who enter for the A-level examination in any subject.

2) There is further evidence of this kind in physics. Massey (1978) found that, for 283 pupils, the correlation between O-level mark and total A-level mark was 0.73. He concluded that 'this may be taken as evidence of validity for the O-level examination'.

3) N.F.E.R. Research Test SF7, Attitude to School.

4) N.F.E.R. Research Test SF4, Free Writing: a 'Torrance type' creativity test.

5) See for example the work of Rosenthal and Jacobson (1968). Their methodology has been strongly criticised (Elashoff and Snow, 1971) but later studies, summarised by Rogers (1982) have confirmed that although the self-fulfilling prophecy effect may not always happen, it certainly can happen. The effect is also discussed at length by Hargreaves (1972).

6) In the light of the conclusions from Year One it was possible to allow schools to choose the activity in which they wished to take part.

7) With the constraints imposed by the manpower available to the project, it was possible to do intensive classroom observation in only one subject - mathematics.

CHAPTER THREE

The Agreement Between Test and Teacher Nomination of More Able Pupils

3.1 Introduction

In Chapter Two we have discussed the derivation of the background criterion which we used to assess the effectiveness of teacher nomination. This criterion was a test based prediction of the pupils' future O-level mark. A simple minded approach to the problem of how to assess the effectiveness of teacher nomination for the top 10% group would be to rank order the pupils in each subject on the basis of this test prediction for O-level success, to mark off the top 10% group on this rank ordered list, to ask teachers to nominate the top 10% in that subject and to count up the number of pupils who were nominated by both the test procedure and by the teachers. This approach would be similar to that adopted in the other studies of this kind which were mentioned in Chapter One; though, as we have already mentioned, the test criterion used in these studies was usually based on measures of IQ. It is quite apparent, however, that to adopt such a procedure we would be to take no account of test error and could, therefore, give misleading results. As a first step, however, we have made, and will report details of, this simple comparison. This is for two reasons. First we will be able, as a result, to make direct comparison between our findings and those of the other workers who have reported their results in this simple way. This is of interest as the earlier studies were concerned with how well teachers could identify children of high general ability. The direct comparison therefore sheds light on the question of whether teachers' ability to identify within a subject differs from their ability to identify general ability. Secondly, we can conjecture that, in the absence of other means of identification, teachers may well use a rank ordered test list in order to find the best pupils for, say, a provision programme. Even if they had reservations in mind concerning test error, in the absence of other information about the pupils, they could do no more in practice than select those who scored highest on the test. It is useful, then, to investigate how close the agreement is between such an unsophisticated test nomination procedure, and teacher nomination.

We will follow the discussion outlined above by what we consider to be a more appropriate discussion of the relationship between teacher nomination and test prediction when test error is taken into account. It has been possible to consider this relationship with regard to the top 10%, top 5%, and top 2%, and also to explore the effect, if any, of the size and type of school.

3.2 Brief summary of the conclusions

a) Teachers were found to be more accurate at identifying pupils of high ability in a subject specific way than had been suugested by the work of previous researchers who were concerned with identification of general (IQ based) ability. This applied to all four subjects, mathematics, physics, English and French, but the effect was particularly marked for mathematics and English.

b) When allowance was made for error of the test, a certain number of mismatches[1] between teacher based and test based predictions could be expected without there being any implication that the teachers had definitely made errors. With these overall numbers of expected mismatches in mind, identification in mathematics and English was found to be very successful. In physics and French, there were larger numbers of mismatches than could be expected, so although there were encouraging signs that the accuracy of identification by teachers was higher than had previously been thought, there was nevertheless some evidence of inaccuracy in these two subjects.

c) There were no large differences in the results for the different types of school in the study. In particular the schools catering for the 13+ age of entry showed similar levels of agreement between test and teacher to the 11+ age of entry schools. Nor were there meaningful differences between the mixed and the single sex schools.

d) The ability of teachers to identify the top 5% and top 2% as discrete groups was not good in any subject, but the percentage of these pupils who were included in the teachers' top 10% lists was encouragingly high.

3.3 Preliminary investigation of the match between test and teacher based identification

Initial results

In Appendix 1 we described a procedure for establishing test-based 'top 10%' groups of pupils in each subject area in each school. We have also drawn attention to the error associated with the tests and the consequent dangers of assuming that the pupils in these test based groups are exactly those who should appear on the teachers' list of 'top 10%' pupils. We have also pointed out that some useful purpose is served in investigating the extent of agreement between these crude test identified groups and teachers' lists. In particlar, such investigations are broadly in line with the methodology used by Pegnato and Birch (1959) and by Solomon (1979) and so allow us to relate our work to theirs. However, it cannot be too strongly stressed that the use of the test results in this simple way disregards the test error. This is a severe limitation of which the reader must be aware.

Before presenting these results it is perhaps useful to discuss the question that we raised in Chapter Two, namely: 'Were the target groups being identified by the tests and the teachers actually the same groups?'. We comment first on what the teachers were asked to do. The form that was sent to each Head of Department is shown in Appendix 2. From it one can see that the teachers were asked to nominate their able pupils in a subject specific way with regard to perceived potential for future performance in the subject. There is no doubt that O-level was largely what the teacher had in mind when assessing how a pupil will perform in the future in his subject. This important point was discussed with the teachers who contributed to the work and we are clear that what the teachers were doing was to pick that 10% whom they felt had the greatest potential for performance at O-level.

The test criterion was also concerned with potential for O-level performance in a subject. It is also clear from Appendix 2 and from the way in which we used the test results (rank ordering pupils within a particular school) that test and teacher were engaged in identification relative to the norms of each individual school. Therefore, we would suggest that the test and the teachers were trying to do the same thing and that, in consequence, comparisons between the two lists of nominees for the top 10% group were entirely appropriate.

later to be used as experimental and control schools, were then made in terms of the effectiveness of their teachers to identify the more able pupils in the Autumn round of identification. This comparison allowed us to establish the extent to which the two groups of schools were equivalent in terms of teachers' abilties to make such identifications. Comparisons between the same two groups of schools were also made in respect of the Spring round of identification to establish whether the checklists, used in the experimental schools in this round, had contributed to any change in the effectiveness of the teachers in those schools. The use of a group of control schools enabled us to separate any effect due to the use of checklists from any effect which might have arisen simply through the greater length of time that the teachers had had to get to know their pupils.

The intention was to seek co-operation from a proportional stratified sample of schools, the total number being sufficient to permit some generalisations of the results to be made, but not so large as to be unmanageable with regard to manpower available for the study. Tables 2.2, 2.3, 2.4 and 2.5 show the numbers of schools in the sample and the numbers of schools in the County of the different types. The only complication to this is where we have one extra single sex 13-18 school which slightly increases the proportion of single sex and 13-18 schools in the sample. This was done, however, so that a comparison could be made between single sex boys' and single sex girls' schools.

The randomization procedure for allocating checklist aided identification to the schools was as follows. Of the three 13-18 schools one of the two single sex schools was randomly assigned to checklist aid in year one. The other single sex and the mixed school were without. This would allow for the possiblity of a further comparison in Year Two when the mixed school could be assigned to the experimental (checklist aided) group.

Of the other eight schools two of the four with 1000+ pupils were randomly assigned to checklists, two of the three with 1000- pupils in the 11-18 range were randomly assigned checklists. The one school in the 11-16 range would not have checklists in the first year and would have them in the second year.

The following tables show the numbers of schools in the county and numbers of schools in the sample of the various types.

TABLE 3.2

OVERALL PERCENTAGES OF TEST IDENTIFIED PUPILS ALSO IDENTIFIED BY TEACHERS

	ENGLISH	FRENCH	PHYSICS	MATHS
ALL SCHOOLS	61 (213)	51 (210)	45 (216)	61 (213)
Range of %	33 - 75	25 - 73	17 - 75	36 - 78

The fact that the school which showed the closest match in English was also the school which showed the worst match in physics is a hint that the range of match between teacher and test judgement is a reflection of teacher variation rather than of the impact of different school size, type or organisation. This hint is confirmed in some of the later work reported in this chapter.

General discussion

It is interesting to compare these figures with that discovered by Pegnato and Birch in their work with 12-14 year old pupils in the junior division of a Junior-Senior High School in Pittsburgh, U.S.A. They asked teachers to nominate children who were 'mentally gifted' and matched these nominations against those who scored 136 or more on an individually administered Stanford Binet IQ test. They found that teachers identified 45% of those who were above this IQ threshold. It is clear that our overall figures were in broad agreement with this result but that they showed rather higher degrees of match between teacher and test, especially in the subject areas of English and mathematics. This higher level of agreement in our experiment may, in part, be a result of our subject specific approach to identification. By choosing to operate in a subject specific way we were asking teachers to engage in a task in which they were interested, and one for which they felt they had appropriate information. Our teachers would not have felt interested in, or well provided with information for the task of assessing general ability (which was the task required of the teachers in Pegnato and Birch's study). The higher level of agreement may also reflect the greater degree of match between what was asked of the teachers and what was asked of the tests in our study. Pegnato and Birch, although they had a very specific test criterion (IQ 136+), simply asked teachers to identify 'gifted children'. The lack of match which they discovered between the nominations produced by teachers and tests may therefore be, at least in part, a consequence of the fact that their teachers may have been thinking in terms of quite different definitions of giftedness than IQ 136+.

We have already pointed out that judgements of teacher accuracy can only be made if due account is made of test error. The comparisons that are being made in this section must be considered with this in mind and what may be suggested from the findings must be investigated in greater depth in later sections. We think, however, that some useful points are worthy of discussion here and it will be seen that these points are not in contradiction with those to be found in the later more detailed study.

In their own right our results provide two interesting insights. First that identification in English and mathematics showed greater match between teachers and tests than was the case in physics and French. This may of course reflect greater inaccuracy in the physics and French test-based identification. However, the evidence provided by the statistics associated with the regression equations which generated the test identified groups (especially the percentage variance explained - 36% in English, 49% in French, 44% in physics and 33% in mathematics - and the standard error of the predicted score - 15 in English, 19 in French, 23 in physics and 30 in mathematics) did not support this view. The equations in physics and French explained more variance in O-level score (i.e. gave predicted scores which were more highly correlated with actual O-level scores) than did the equation in mathematics and also resulted in lower standard errors. The physics and French equations also explained more variance than the English equation though admittedly, in this case, the standard errors were higher.

It has been possible to do a small scale study of the correlation between the French predictor and another predictor for achievement in French, and also to compare prediction with actual O-level score in mathematics in one school. Both of these give encouraging indications and there is a fuller discussion in Appendix 1. There seems no reason to suppose, therefore, that the French and physics equations are any less successful than the English and mathematics equations at predicting O-level score. However, the lower agreement between teachers and tests in physics and French may reflect the fact that, in these subjects O-level score is, in itself, an inappropriate criterion measure with which to compare teacher judgements.

On the forms that were issued to teachers, they were asked to identify 'that 10% of third year pupils whom you expect to have the greatest potential for further work in your subject' and it could be argued that, in French and physics, potential for O-level performance does not coincide with this specification. While we would stress the point made earlier in this chapter (namely that, in discussion with the teachers it was clear that O-level potential was firmly in their minds) it could be that physics and French teachers tried to take account of other aspects of high ability. This difference in emphasis could of course contribute to the greater mismatch between tests and teachers in these subjects. We feel that the force of this argument is attenuated by the fact that in English, the subject, where that which constitutes potential for further work in the subject might be expected to be furthest removed from that which leads to success in O-level English Language, the teachers and tests were in substantial agreement. We feel too that, in physics in particular, the nature of the subject makes it difficult to sustain the view that a third year pupil could have potential for further work without having potential for O-level success. We should also, perhaps remind the reader of the relevant work which was quoted in Chapter Two where we reported that high correlations between O-level and A-level performance in physics had been established by Murphy (1981) and by Massey (1978). This would suggest that, even if teachers were looking beyond O-level, the pupils they picked should still be the high O-level scorers.

It would seem therefore that the greater mismatch between test and teacher judgements in French and physics may indicate a genuinely greater difficulty for teachers in making judgements about pupils in these subjects. We suggest that this view is supported by the work on the characteristics of the teacher selected groups which is reported later. Reasons for this difficulty can only be speculative at this stage but may include:

i) The difficulty of making accurate assessments of individuals' work in class (oral work in French, practical work in physics);

ii) the nature of the syllabus content in the first years (e.g. in physics little challenging mathematical work is undertaken below Year 4 so that pupils have little opportunity to display ability of this type);

iii) the small amount of teacher contact with classes (especially in physics where only an hour or so is spent in each class each week). This difficulty is compounded by the consequently large number of classes that the teacher has to try to get to know;

iv) the fact that, at the beginning of the third year, the work in physics (and to a lesser extent French) is at a fairly simple level because the subject is fairly new to the pupil. In mathematics and English, which have five primary years to build on, the work is beginning to get more challenging, and is more likely to provide opportunities through which the more able can demonstrate their ability.

School differences

Differences between individual schools

Table 3.1 indicates that, within each subject, schools seemed to vary in the extent to which teachers' and tests' judgements agree. To explore the question of whether or not these differences were too large to be attributable to chance, the data was cast into the form shown in Table 3.3 and analysed by means of the chi-squared test.

TABLE 3.3

TEACHER/TEST AGREEMENT: DIFFERENCES BETWEEN SCHOOLS IN ENGLISH

SCHOOL	Test identified pupils	
	identified by teacher	missed by teacher
00	8	6
01	12	6
02	10	9
03	10	5
04	12	6
05	9	7
06	14	6
07	11	11
08	9	3
09	17	7
13	18	17

X^2 = 6.60

df = 10

not significantly different (nsd)

Similar analyses in the other subject areas are reported in Table 3.4 and lead to the conclusion that there were no statistically significant differences between individual schools in the numbers of test identified pupils who were also identified by teachers.

TABLE 3.4

TEACHER/TEST AGREEMENT: SUMMARY OF DIFFERENCES BETWEEN SCHOOLS

	ENGLISH	FRENCH	PHYSICS	MATHS
X^2	6.60	10.17	12.14	7.95
df	10	10	10	10
Significance	nsd	nsd	nsd	nsd

Differences between 11+ entry and 13+ entry schools

Schools were grouped according to whether they accepted pupils at age 11 or age 13. Tables 3.5 and 3.6 show the results of comparisons between these groups.

TABLE 3.5

COMPARISON OF TEACHER-TEST AGREEMENT BETWEEN 11+ ENTRY AND 13+ ENTRY SCHOOLS IN ENGLISH

	Test identified pupils		% of test identified pupils identified by teachers
	identified by teachers	not identified by teachers	
11+ entry	94	60	61
13+ entry	36	23	61

$X^2 = 0.024$

df = 1

nsd

TABLE 3.6

SUMMARY OF DIFFERENCES BETWEEN 11+ ENTRY AND 13+ ENTRY SCHOOLS

Subject	X^2 (df=1)		% of tests identified pupils also identified by teachers in schools with	
			11+ entry	13+ entry
ENGLISH	0.02	nsd	61	61
FRENCH	2.79	nsd	47	61
PHYSICS	2.18	nsd	41	53
MATHS	2.17	nsd	57	69

The chi-squared values shown in Table 3.6 were calculated in the same way as that shown in Table 3.5 and were corrected for continuity (Seigel, 1956). The percentage figures are quoted to give an indication of the nature of the differences which were under investigation in the chi-squared analysis. The overall, and perhaps surprising conclusion, is that there were no statistically significant differences between the 11+ entry and 13+ entry schools. Perhaps more surprising is the fact that the trend indicated by the results was that the 13+ entry schools showed greatest agreement between test and teacher judgement. It has been suggested to us by several heads of 11+ entry schools, that this result may reflect the fact that there is often a substantial change of staffing between 2nd and 3rd year classes so that the additional opportunities for staff to get to know pupils in an 11+ entry school are more illusory than real.

An alternative suggestion is that, in 13+ entry schools, the first term of the school year is naturally a time when staff are attempting to assess the new 3rd year intake whereas in 11+ entry schools assessment of this year group at this stage of the year is a less pressing priority. It may be that the lower degree of match between teacher and test judgement in 11+ schools was a consequence of this fact: namely that the exercise was, for the staff of these schools, more artificial.

Effects of school size

For this analysis the 11+ entry schools were grouped according to size. School 13 formed a 'group' of its own with 2000+ pupils. Schools 03 and 08 formed a group of 'small schools' with less than 1000 pupils. Schools 00, 01, 04, 06 and 07 formed a group of middle sized schools (1000 to 1500 pupils).

The analysis of school size difference was conducted in a similar fashion to that reported in the preceding section and the summary table is given below.

TABLE 3.7

SUMMARY TABLE OF SCHOOL SIZE DIFFERENCES

Subject	X^2 value (df = 2)		% of test identified pupils also identified by teachers		
			large	medium	small
ENGLISH	2.38	nsd	51	62	70
FRENCH	0.11	nsd	47	46	50
PHYSICS	2.68	nsd	53	38	34
MATHS	1.41	nsd	59	58	46

No statistically significant differences were detected and no clear trends were discernable in the data from this aspect of the work.

These results taken together provide support for the hint to which we drew attention earlier: namely that variation was not dependant on gross organisational matters such as size or type of school but that it probably reflected variation in the effectiveness of individual teachers.

3.4 Comparison of teacher nomination with test prediction when test error is taken into account

We now move on to what we feel is a more important aspect of the analysis, in which we have attempted to take account of test error when comparing the teacher generated nominations for the top 10% group with those arising from the tests. We have made this more meaningful comparison in two different ways.

We will first describe the simpler approach in which the rank ordered lists were divided into three groups: Group A containing those pupils who, even with test error in mind, scored high enough to rank them as pupils who were very strong candidates for membership of the top 10% group, Group B containing those pupils, all of whom scored high enough to be reasonable choices for the top 10% group and Group C containing the remainder of pupils, all of whom scored so low that even with test error in mind the pupils were weak candidates for inclusion in the 10% group. We will discuss the relationship of the teacher's identification to these three groups. In the second comparison, we will use a knowledge of the magnitude of the test error to calculate the total numbers of pupils, in various bands of test score, whom we could expect, because of test error alone, to be identified by teacher and not by test, or by test and not by teacher. We can, in this way, judge if the level of mismatch is beyond reasonable expectation for each subject. If this proves to be the case then we can argue that there is evidence for error on the part of the teacher as well as the test.

The A, B, C Groups

The first exercise was to rank order the pupils for each subject and each school according to their predicted O-level score from the regression equation. A pupil who was missing for any of the relevant background tests was excluded as a 'missing data' case whether or not he was on the teacher's list in that subject. Though the teacher was asked to nominate the top 10% there was some fluctuation in the numbers of pupils appearing on the lists up or down from the true 10% of the year group. In order to draw comparisons with test prediction, therefore, a line was drawn across the test ranked list so that it included above the line the number of teacher nominations, reduced by the number of missing data cases on the teacher's list. Simple comparisons of numbers of pupils identified by teacher and test could then be tackled. A typical computer listing would look like Table 3.8.

TABLE 3.8

TO ILLUSTRATE THE USE OF THE PUPIL'S LIST RANK ORDERED BY THE TESTS

	SCHOOL 20	SUBJECT ENG		
	NAME	PRED. SCORE	CODE*	
A GROUP	pupil 1	125.2	1	
	pupil 2	123.9	0	
	pupil 3	122.6	1	
B GROUP	pupil 4	120.2	1	
	pupil 5	119.3	0	
	pupil 6	119.1	1	
	pupil 7	116.2	0	
	pupil 8	114.7	0	
	pupil 9	113.5	1	
	pupil 10	113.2	0	
	pupil 11	112.9	0	
- - - - -	- - - - -	- - - - -	- - - - -	10% LINE
	pupil 12	112.7	1	
	pupil 13	111.8	0	
	. .	.	.	
	. .	.	.	
	. .	.	.	
	. .	.	.	
	. .	.	.	
	pupil 21	108.6	1	
C GROUP	pupil 22	106.3	0	
	pupil 23	105.1	0	
	pupil 24	105.0	1	
	. .	.	.	
	. .	.	.	
	. .	.	.	
	. .	.	.	
	. .	.	.	
	pupil 93	63.4	0	
	pupil 94	67.8	0	
	. .	.	.	
	. .	.	.	
	. .	.	.	

---

* Code is 0 if not on teacher's list, 1 if identified by teacher.

As we have already stated, we were not content with just a simple count of right or wrong above the 10% line for several reasons. If one looks at the predicted scores of the pupils either side of the 10% line one sees only a small difference in scores. If one compares this difference, and indeed differences in scores for pupils even further apart, with the known test error (the standard error is shown with the regression equations) one sees that there is a high probability that when these pupils do their O-level they will have a different rank ordering from that in the predicted list. By knowing the magnitude of the test error we can make allowances in our analysis.

We treat each person's test score as his predicted O-level score and imply that his actual O-level score will lie within a normal distribution, centred on this predicted score and having a standard deviation equal to the standard error from the prediction equation[2]. We then divide the rank ordered list into three groups, as shown in Table 3.8: Groups A, B and C. Group 'A' is the set of pupils who, even when taking test error into account, have a high probability of performing on actual O-level so that they remain within the top 10%. The lowest mark for the A group is the sum of the mark that defines the 10% boundary (note that this mark is relative to the school) added to a figure equal to (0.675 x standard error). With the A Group defined in this way, there is a probability of at least 0.75 that a pupil who has a mark above this minimum will be in the top 10% when he/she attempts O-level. Thus the 'A' group represents the pupils with high potential with regard to test assessments even when test error is taken into account. When teachers overlook several members of the A Group in making up their own nominations, one can begin to question the accuracy of their judgement.

The 'B' Group includes pupils whose scores were within ±0.675 x Standard Error of the mark which defines the 10% line. They represent the pupils all of whom have, in the light of their own predicted score and of the test error, a good chance of being in the top 10% on O-level. Selection of a B Group pupil by the teacher therefore counts as a perfectly reasonable choice on that teacher's part. Some of these B Group pupils are below the 10% line on predicted score and some are above. The ones above the line would be reckoned to have the best chance of success, but of all of the pupils who gained the same predicted score as any one B Group pupil, at least 25% in every case would achieve an O-level score above the mark which defines the top 10% group (based on our knowledge of the prediction equation and its error). Thus in a sense the B Group is an homogeneous group in that all of its members are reasonable candidates for membership of the top 10% group in terms of actual O-level performance. Group B is also homogeneous in a second sense. The two pupils who scored the top mark and the bottom mark in this group have at least a 0.17 probability of reversing their positions at O-level[3] and thus of course, pupils closer on predicted score have higher probabilities of changing rank orderings in the O-level exam.

Group C consists of the rest of the year group. Even the best pupils in this group have only a probability of 0.25 that their actual score at O-level will bring them into the top 10% group. For many of these pupils who score well below the boundary score for the group, the probability of a good O-level score is very much less than 0.25.

The choice of the 75% probability level for Group 'A' is arbitrary in one sense but was designed to allow Group 'B' to be a homogeneous group as far as possible. The fact that only a small number of pupils appear in Group A is an indication that this approach to the interpretation of the test scores produces a model which closely simulates the situation which our experience leads us to expect when engaged in any identification process. There is usually a small group of people who are clearly members of the target group, a further group who are clearly not members and a group about whom we cannot be certain. Group 'A' represents the outstanding pupils for whom there is little doubt (in this case little doubt as measured by test scores). Group 'B' represents a group from which a choice must be made but in this choice there is some uncertainty. Any teacher selection in the B Group is a reasonable selection and the characteristics of the chosen pupils, compared with the others, can be taken as indicating the characteristics that teachers respond to when a choice has to be made.

Thus the division into the three groups is the basis of our attempt to gain deeper insights into the identification process that cannot be found from simple counting exercises based on the 10% line alone.

3.5 The match between teacher and test judgements in terms of groups A, B and C

If we make allowance for the test error, we are admitting to the inaccuracy of the test in predicting the potential of children for future O-level performance. However, almost paradoxically, once we have begun to take account of the test error we can also begin to interpret differences between test and teacher based nominations not just as signs of disagreement between these two methods of identification, but actually as signs of error on the part of the teacher. In the simple work reported in Table 3.1 we knew that there was some disagreement between test and teacher nomination. As we knew that there was error in the test we could not imply that this disagreement was necessarily a sign of teacher error. However as we begin to quantify the test error we can begin to make comments on the accuracy of teachers' judgements. For example, we know that even in the light of test error, at least 75% of the pupils in the A Group will score well enough on O-level to be members of the top 10% (though we are not able to tell from the tests which of the A Group pupils will achieve this). If therefore teachers overlook say 60% of the A Group, then there is not only evidence of disagreement between teacher and test, but also fairly firm evidence that the teacher has made an error.

It may be useful for the reader if we repeat the interpretation of the A, B, C groups and give also, at this point, the definition of the other sort of classification of pupils which we have used in following sections. (This alternative classification is as 'Good', 'Satisfactory' and 'Unlikely' members of the 10% group). The relationship between these two classifications is very simple as will be apparent from Table 3.9. Both types of classification are used later as each is particularly appropriate for some kinds of analysis.

GROUP A	those who had a 75% or more chance that their actual O-level score would confirm their membership of the 10% group;
GROUP C	those who had a 75% or more chance that their actual O-level score would confirm that they should not be in the 10% group;
GROUP B	those in between. Despite the fact that, on the prediction of O-level score some of this group fell in, and some out, of the 10% group the pupils in the B group were fairly homogeneous (there was at least a 17% chance that any two could change positions) and any pupil in this group could be thought of as a

reasonable candidate for inclusion in the 10% group.

TABLE 3.9

		) G		
	Pupil 1	) R		75% + chance of being
		) O	A	'above the 10% line'
Good candidates	Pupil 2	) U		
for membership		) P		
of 10% group		)		
		(		
	Pupil 3	(		
		(		
	Pupil 4	(G		
		(R		
	Pupil 5	(O	B	
- - - - - - - - - - - - - - - -		-(U		- - - -10% group borderline
		(P		
Satisfactory	Pupil 6	(		
candidates		(		
	Pupil 7	(		
		)		
	Pupil 8	) G		
		) R		
Unlikely	Pupil 9	) O	C	
candidates		) U		75% chance of being
	"	) P		'below the 10% line'
	"	)		
	"	)		

Of course, although Pupils 3 to 7 form the B group, Pupil 3 did achieve a higher predicted score than Pupil 7 so was a rather better candidate for membership of the 10% group. Viewed in this way the alternative grouping of pupils can be established. In terms of Table 3.9, Pupils 1 to 5 were the 'good' candidates for inclusion in the 10% group. Pupils 6 and 7 were 'satisfactory' candidates and pupils 8 and 9 etc. were 'unlikely' candidates. The spread of teacher nominations across these three groups will now be considered.

TABLE 3.10

PERCENTAGE OF TEACHER NOMINATIONS FALLING IN THE THREE GROUPS: 'GOOD', 'SATISFACTORY' AND 'UNLIKELY'

SUBJECT	n	Percentage of teacher nominations who were: Good Candidates	Satisfactory Candidates	Unlikely Candidates
ENGLISH	213	61	27	12
FRENCH	210	51	13	36
PHYSICS	216	45	25	30
MATHS	213	61	23	16

The figures in the final column of this table reinforce the view that identification in French and physics showed the greatest mismatch between teacher and test judgements. It seems that 36% of teacher nominations in French, and 30% in physics, were pupils with such low predicted scores that, even given the benefit of the doubt which was generated by the test error, these pupils were very unlikely to get high enough O-level scores to bring them into the top 10% group. Of course these figures can be presented in an alternative way: if 30% of physics nominations are, in the sense set out above, likely errors, 70% can be thought of as perfectly reasonable nominations. In a similar way we can report that 88% of nominations in English, 64% in French and 84% in mathematics were reasonable nominations.

Reported in this way the situation looks rather different from, and more encouraging than, that inferred from the work of Pegnato and Birch. However, two important questions remain. First, what was it about the unlikely pupils which led teachers to nominate them as members of the 10% group? Secondly, what is it about the reasonable nominations which were made, which set them apart, in teachers' minds, from the other equally reasonable pupils who were not nominated? This important point will be considered in the next chapter and in those later chapters where we report on the work done during the second year of the project. Before looking more closely at the A and C group we turn to the question of whether school size, or the age of the main intake, affected the distribution of teachers' nominations across the three categories (good, satisfactory and unlikely candidates for 10% group membership). Table 3.12 gives the full table of figures for all schools for physics which was subjected to the chi-squared test. Table 3.11 summarises the results for all four subjects which were analysed in the same way.

TABLE 3.11

ANALYSIS OF INDIVIDUAL SCHOOL DIFFERENCES (SECTIONS OF SCHOOL 13 ANALYSED SEPARATELY)

Subject	Chi-squared values df = 24	
ENGLISH	20.2	nsd
FRENCH	31.4	nsd
PHYSICS	43.1	1%
MATHS	16.5	nsd

TABLE 3.12

PHYSICS NOMINATIONS: SCHOOL DIFFERENCES

	Number of nominations which were of		
School	Good Candidates	Satisfactory Candidates	Unlikely Candidates
00	7	5	4
01	9	8	4
02	11	2	8
03	7	3	4
04	4	5	11
05	8	2	5
06	11	6	6
07	6	2	8
08	2	7	3
09	13	7	2
10	9	0	3
11	2	6	2
12	7	1	4

In physics but in no other subject a statistically significant difference was found between the schools treated individually[4]. It seems from the chi-squared table that, in physics, school 04 nominated more unlikely pupils than expected whereas school 09 nominated more good and satisfactory pupils than expected. No explanations come to mind for this difference other than the simple one: that teachers vary in their ability to identify able children.

In all four subjects we found that there were no statistically significant differences between the 11+ entry and 13+ entry schools. Within the 11+ entry group schools of different size again showed no significant difference in the way their nominations were distributed. These results are summarised in Tables 3.13 and 3.14.

TABLE 3.13

DIFFERENCES BETWEEN 11+ ENTRY AND 13+ ENTRY SCHOOLS

Subject	Chi-squared values df = 4	
ENGLISH	0.00	nsd
FRENCH	4.07	nsd
PHYSICS	3.56	nsd
MATHS	2.26	nsd

TABLE 3.14

DIFFERENCES WITHIN THE 11+ ENTRY GROUP ANALYSED IN TERMS OF SCHOOL SIZE

Subject	Chi-squared values df = 4	
ENGLISH	3.29	nsd
FRENCH	1.02	nsd
PHYSICS	4.66	nsd
MATHS	0.45	nsd

3.6 The level of mismatch in groups A and C

We recall that the A Group contained those pupils who, even when test error was taken into account, are likely to have an O-level score sufficiently high to confirm their membership of the top 10% and the C Group contained pupils who in the same sense were unlikely to obtain membership of the top 10%. We can, therefore, consider the level of mismatch between test and teacher nomination within these two groups and make some preliminary judgements about the accuracy of the teachers' judgements. If teachers were making completely accurate identifications we would expect few mismatches in these groups.

Table 3.15 shows, for all schools together, the percentage of pupils in the A Group who were nominated by teachers.

TABLE 3.15

PERCENTAGE OF A GROUP PUPILS NOMINATED BY TEACHERS

	ENGLISH	FRENCH	PHYSICS	MATHS
ALL SCHOOLS	80.8	61.4	53.7	76.2

Analysis of the data which generated the results in Table 3.15 was undertaken to establish whether there were differences between individual schools, between schools with 11+ and 13+ intakes, between schools of different size and between the experimental and control schools. The analysis to investigate individual school differences, and that to investigate the effect of size were hampered by the fact that very small numbers were involved. However the results of these analyses (treated with caution because of the problem of small numbers) did appear to indicate that there were no significant differences.

The analyses of schools grouped by age of entry did not suffer from small number problems and in Table 3.15a we quote the results of ths work. In common with earlier results of a similar kind, no significant differences were detected.

TABLE 3.15a

DIFFERENCES BETWEEN GROUPS OF SCHOOLS IN THE EXTENT TO WHICH TEACHERS NOMINATED A GROUP PUPILS

	ENGLISH	FRENCH	PHYSICS	MATHS
X^2 value 11+ entry schools	0.00	3.48	1.03	0.04
against				
13+ entry schools (df = 1)	nsd	nsd	nsd	nsd

It is perhaps surprising that after half a term's knowledge of the pupils, teachers in all schools should overlook between one-fifth (in English) and almost one-half (in physics) of those pupils who, on the tests, fell into this high scoring 'A Group' category. In view of the definition of the A Group, while the English result does not force us to the assumption that errors were made by the teachers, the physics result does stongly suggest that this was the case.

Table 3.16 shows, for all schools together, the percentage of pupils nominated by teachers who were in the C Group.

TABLE 3.16

PERCENTAGE OF TEACHER SELECTED PUPILS WHO WERE IN THE C GROUP

	ENGLISH	FRENCH	PHYSICS	MATHS
ALL SCHOOLS	12.2	35.7	30.6	15.8

As in the A Group analysis we found no indication of individual school differences in the extent to which teachers' nominations fell into the C Groups. We also found no indication of differences between schools grouped according to size. Again, as in the A Group analysis, small number problems rendered the chi-squared tests used to establish these results rather suspect. However, the analysis of schools grouped according to age of intake avoided small number problems and again failed to disclose any significant differences.

Consideration of the results indicated to us the greater accuracy in English and mathematics compared with French and physics and led us to the analysis that will now be reported and then to a search for explanations for these mismatches, that will occupy the next chapter of the report. Though there are good signs that teachers are more accurate at subject specific identification than previous research has shown for identification based on general ability, there is sufficient inaccuracy to cause concern and a need to search for explanations.

3.7 The mismatch between test and teacher identification when fuller allowance is made for the test error

The creation of Groups A, B and C was the basis of an investigation which made some allowance for error in the background test predictions. By having an estimate of this error we could work, within known limits of the test accuracy, to draw conclusions about the accuracy of teacher identifications.

Within each school and for each subject it would not have been possible to subdivide the year group into more than three groups because of the small numbers that would have been generated. Indeed the groups A, B and C that did emerge were rather similar to the kinds of groups a teacher might have in mind: the small number (Group A) who gave little doubt of their ability, a larger group, such as the rest of a top set in the subject, each of whom was a reasonable choice (Group B) and the rest of the year group (Group C) all of whom were rather weaker candidates for the top 10%.

As we have been able to show that there were few differences between schools through our analysis in terms of Groups A, B and C, we decided that it would be helpful to carry out a more detailed analysis by pooling the results from all of the schools. This more detailed analysis is essentially based on division of the test based rank order list of pupils into smaller bands, and it is because of this finer subdivision that the pooling of results was necessary so that small number problems could be minimised. Had we discovered major differences between individual schools, the validity of this investigation on the accumulated data would have been far more open to criticism but, as we have indicated, this particular problem does not arise on this occasion.

Rationale

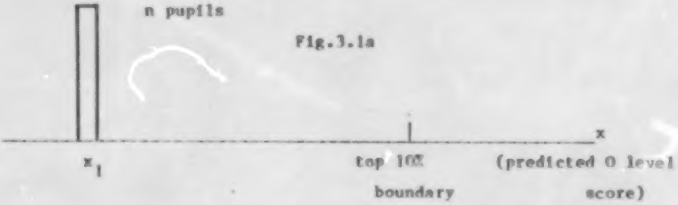


Fig.3.1a

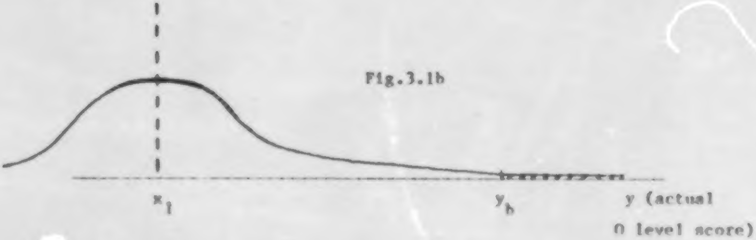


Fig.3.1b

Suppose we look at a group of n pupils who all have predicted score within a small band centred on x_1 (Fig. 3.1a). If we find that a teacher chooses n_t of these as members of the top 10% we might conclude that the teacher must have made n_t 'false hits' (for x_1 is well below the 'top 10% boundary score'). In fact, because of the error associated with the test, we cannot be certain that these n pupils will actually get an O-level score of x_1. Indeed we can be almost certain that they will not all get this score. Their actual O-level scores will be as shown in Fig. 3.1b. They will follow a normal distribution centred on x_1 with a standard deviation equal to the standard error of the test. As shown, a consequence of this distribution is that a number of the n pupils with predicted score in a small band centred on x_1, will gain an actual O-level score which is above the top 10% boundary, y_b. The number to do so will be proportional to the shaded area in Fig. 3.1b. If this number is less than n_t then we can continue to say that the teacher has made some errors. (This was found to be the situation in physics and French). We can see however that if (as proved to be the case in English and mathematics) this number is equal to n_t then our original assumption that the teacher must have made some errors cannot be substantiated. The teacher may have picked the very pupils who will do sufficiently well 'on the day' to justify his faith in them in their 3rd year. On the other hand it may be that the teacher's nominations from this group are not the ones who actually score highly later on. Without actual O-level results to refer to we can only attempt to distinguish between these possible interpretations by looking at the kind of pupils who make up the teachers' 'false hits'. (If, for example, physics 'false hit' pupils had high Numerical Ability scores we might be inclined to regard them as appropriate nominations. If the only thing which distinguished them was that they were neat, we might conclude that the teacher was probably being misled). The study of these characteristics is the subject of the next chapter. We deal here with the accuracy of nomination in quantitative terms.

Analysis

For each school and subject we counted up how many pupils appeared in bands, each half a standard error wide from the mark which defined the 90th percentile on predicted score[5]. The total number was then found in each subject and band for all schools. This is shown in Table 3.17.

TABLE 3.17

TOTAL NO. OF PUPILS WHOSE PREDICTED SCORE LAY IN A BAND OF WIDTH ½ A STANDARD ERROR, BASED ON A DATUM AT THE 90th PERCENTILE

	BELOW 90th PERCENTILE						ABOVE 90th PERCENTILE				
	2½-3	2-2½	1½-2	1-1½	½-1	0-½	0-½	½-1	1-1½	1½-2	2-2½
ENGLISH	-	190	337	318	256	179	102	64	39	6	-
FRENCH	169	191	186	160	116	89	84	53	32	25	12
PHYSICS	145	305	379	400	319	206	121	55	24	9	2
MATHS	-	265	355	290	242	159	113	62	23	13	-

In each of these bands we then counted up the apparently 'mismatched' pupils i.e. those pupils above the 90th percentile on score who were not on the teachers' lists and those pupils below the 90th percentile on score who were on the teachers' lists. These numbers appear in Table 3.18. We then attempted to estimate the number of pupils from each band below the 90th percentile who could be expected, on the basis of the test error, to be in the top 10% group on actual O-level performance; and the number of pupils in each band above the 90th percentile who could be expected, on the basis of test error, not to be in the top 10% group on actual O-level performance. We assumed that each pupil in each band scored the mean score of that band and that on O-level these pupils' scores would be normally distributed about this mean score with a standard deviation equal to the standard error. On this assumption it was possible to estimate the position of the new ten per cent line for O-level and to locate how many from each band of predicted score would be in the top ten per cent on O-level, as defined by this line. It was, therefore, a simple task from this to estimate how many of the test predicted top 10% would in fact be wrong predictions and how many of the non test predicted would be in the top 10% at O-level. In addition we would have an estimate of how many of these mismatches would come from each of the bands of predicted scores. These numbers are shown in Table 3.18.

TABLE 3.18

NOS. OF APPARENT INCORRECT IDENTIFICATION FROM TEACHERS' LISTS COMPARED WITH NO. OF FORECAST MISMATCHES ACCORDING TO TEST ERRORS IN THE REGRESSION EQUATIONS. (BANDS ½ S.D. WIDE)

	BELOW 90th PERCENTILE						ABOVE 90th PERCENTILE				
SUBJECT	2½-3	2-2½	1½-2	1-1½	0-1	0-½	0-½	½-1	1-1½	1½-2	2-2½
ENGLISH											
TEACHER	-	0	3	8	27	45	56	19	7	1	0
TEST	-	1	5	15	28	43	58	14	10	0	0
FRENCH											
TEACHER	7	15	13	20	27	18	51	26	12	9	2
TEST	0	1	5	13	21	28	41	17	7	2	0
PHYSICS											
TEACHER	2	8	8	24	38	39	77	26	13	2	1
TEST	0	1	5	16	35	46	71	23	7	2	0
MATHS											
TEACHER	-	3	3	13	29	38	54	23	5	3	1
TEST	-	0	6	15	30	40	62	22	7	1	0

There are two categories of 'mismatch' in Table 3.18. Above the 90th Percentile we are concerned with the number of pupils who we can expect to score high on the 3rd Form test, but will not eventually do so on their O-level. Below the 90th Percentile we are concerned with those who score low on the 3rd Form test but will eventually achieve top 10% marks at O-level. Consider the part of the table which refers to pupils who are below the 90th percentile. In each band we know the number of pupils who could, on the basis of the test, be expected to get into the top 10% (see Table 3.18). We also know how many could be expected not to reach this standard (from the differences between appropriate numbers in Tables 3.17 and 3.18). From Table 3.18 we also know how many pupils from the band in question were regarded as members of the top 10% group by their teachers; and can calculate from these figures and the appropriate figures in Table 3.17, how many were not regarded as being in the top 10% by their teachers. We then used a chi-squared 'Goodness of Fit' test to compare the teachers' figures with the expected figures[6]. The results of these statistical tests are shown in Table 3.19.

TABLE 3.19

CHI-SQUARED VALUES AND SIGNIFICANCIES OF NUMBER OF APPARENTLY INCORRECT IDENTIFICATIONS BY TEACHERS COMPARED WITH NUMBERS OF FORECAST MISMATCHES BASED ON THE TESTS WHICH WOULD TURN OUT TO BE CORRECT NOMINATIONS

		BELOW 90th PERCENTILE				ABOVE 90th PERCENTILE		
ENG/MATH		1½-2½ SD	1-1½	½-1	0-½	0-½	½-1	1-2½ SD
FR/PHYS		1½-3 SD	SD	SD	SD	SD	SD	SD
ENGLISH	X^2	1.52	3.84	0.16	0.12	0.16	2.29	0.51
	SIG	NS	NS	NS	NS	NS	NS	NS
FRENCH	X^2	141.08	4.10	2.09	5.21	4.76	7.01	18.4
	SIG	≤0.1%	5%	NS	5%	5%	1%	≤0.1%
PHYSICS	X^2	24.17	4.17	0.29	1.37	1.23	0.67	7.33
	SIG	≤0.1%	5%	NS	NS	NS	NS	1%
MATHS	X^2	0	0.28	0.04	0.13	2.29	0.07	0.16
	SIG	NS	NS	NS	NS	NS	NS	NS

It is encouraging in both English and mathematics that the number of mismatches corresponds for both teacher and test predictions in all of the bands of scores. We must, of course, repeat that even if numbers correspond the particular pupils may not, but certainly there is evidence that at least some of the listed and apparent mismatches may be genuinely able pupils and only in the light of further information can more be said.

In the case of French, even allowing for test error, there is still a worrying discrepancy between the numbers of pupils who may be expected to score differently from their predicted score and the number of high scores who are missed, or low scorers who are selected by teachers. It would appear that far too many low potential pupils are being assessed among the top 10% and far too many highly able are being missed. In physics there is a similar if not quite so marked discrepancy. It is interesting to note that in both subjects the discrepancy is noticeable among pupils whose score was more than half a standard deviation either side of the 90th percentile score and becomes more acute in the extremes of the score ranges. These areas of worry correspond roughly to the A and C Groups of our main analysis.

We are therefore led to believe that, in physics and French, the numbers of pupils missed from the A Group or selected from the C Group do represent teacher error. In English and mathematics the number of mismatches is in line with what, as a consequence of test error may be considered reasonable.

We are then left with the interesting problem of attempting to establish whether the actual individuals who are judged differently by test and teacher are in fact correctly identified by the teacher. Until actual O-level result are available, this can be done only by looking at the characteristics of the chosen pupils and assessing the appropriateness of these characteristics. Though some insights can be gained by looking at the characteristics which we have already measured and which is the substance of the next chapter the work can only be developed by detailed case studies. This was seen as one priority for the second year of the research.

In the absence of additional information it remains true, of course, that the individuals with the greatest chance of having an actual O-level score which places them in the top 10% group are those with the highest predicted scores. Our analysis in terms of A, B and C Groups is not therefore invalidated by this further work. It simply emphasises the need for further data and provides a way of interpreting such figures as 36% nominations from the C Group as indicating actual teacher error (rather than, as indicating mismatch which we can only suggest may reflect teacher error).

It is interesting to note that this procedure allows us to engage in a very detailed discussion of teacher error despite the fact that there is error in the criterion measure. This contradicts the statement made by Rogers (1982) that:

> If the criterion does not accurately indicate the true (nature of) the object person then it is clearly useless as a criterion measure.

We feel that this is an important contribution to the debate about accuracy measures in research, as Rogers' view does tend to stifle research initiatives as 'accurate' measures are very difficult to find.

Another general point needs to be made here. Although our prediction equations can be effective in monitoring teachers' judgements (when they are used in probabilistic statements about groups of children) they are, as we have continually stressed, relatively poor as the sole tool for use in isolation, in identifying individual pupils. They would need to be subject to far smaller errors if we were to rely on them for the actual process of individual identification. This will be clear to the reader from the whole structure of the argument used in this section, but we feel that it is important enough to bear specific mention at this stage.

Comparison with random nomination

It is often an enlightening exercise to compare a strategy (such as the teacher nomination discussed here) with a strategy based on completely random allocation. Whether or not the strategy is reasonably accurate, if it can do no better than random, then the strategy must be reviewed. Table 3.20 gives the total number of nominations that would come from each band if random allocation to the top 10% were to be used. These figures are computed from Table 3.17. by the simple process of taking 1/10th of the appropriate figure for each cell of the table.

TABLE 3.20

TOTAL NO. OF PUPILS IN BANDS OF WIDTH ½ STANDARD ERROR WHO WOULD BE NOMINATED BY RANDOM ALLOCATION TO THE ABLE GROUP

•	BELOW 90TH PERCENTILE						ABOVE 90TH PERCENTILE				
	2½-3	2-2½	1½-2	1-1½	½-1	0-½	0-½	½-1	1-1½	1½-3	2-2½
ENGLISH	-	19	34	30	26	18	10	6	4	1	0
FRENCH	17	19	19	16	12	9	8	5	3	3	1
PHYSICS	15	31	38	40	32	21	12	6	2	1	0
MATHS	-	27	36	30	24	16	11	6	2	1	0

Firstly we can compare teacher nomination with Random assignment by carrying out chi-squared tests for 'Goodness of Fit' in each of the bands using the Random allocation numbers to generate the expected Table. This is a similar method as that which was used to compile Table 3.19, so that the contingency tables make use of figures from Tables 3.17, 3.18 and 3.20 (see Note 6). Table 3.21 summarises the results.

TABLE 3.21

CHI-SQUARED VALUES AND SIGNIFICANCIES TO COMPARE TEACHER NOMINATION WITH RANDOM ASSIGNMENT OF PUPILS TO THE TOP 10%

		BELOW 90th PERCENTILE				ABOVE 90th PERCENTILE		
ENG/MATH		1½-2½ SD	1-1½	½-1	0-½	0-½	½-1	1-2½ SD
FR/PHYS		1½-3 SD	SD	SD	SD	SD	SD	SD
ENGLISH	X^2	52.44	17.81	0.04	45.03	234.6	31.08	2.03
	SIG	≤.1%	≤.1%	NS	≤.1%	≤.1%	≤.1%	NS
FRENCH	X^2	8.09	1.11	22.48	10.01	255.45	97.39	40.70
	SIG	1%	NS	≤.1%	1%	≤.1%	≤.1%	≤.1%
PHYSICS	X^2	57.70	7.11	1.25	17.91	409.19	74.83	61.61
	SIG	≤.1%	1%	NS	.1%	≤.1%	≤.1%	≤.1%
MATHS	X^2	57.60	10.74	1.16	33.63	186.22	53.33	13.09
	SIG	≤.1%	.1%	NS	≤.1%	≤.1%	≤.1%	≤.1%

In order to draw conclusions about the numbers of pupils picked we must now look at Tables 3.18 to 3.21 in conjunction. Table 3.19 allows us to see where areas of concern exist. Table 3.21 shows if random allocation gave similar or significantly different numbers, in the various bands, to the teacher and a comparison of Teacher, Test and Random allocation then reveals the direction of any discrepancy.

In English and mathematics, in terms of numbers expected there was considerable agreement between test and teacher, but in all but 3 cells of Table 3.21 there would be significantly larger differences if random allocation had been adopted, so we can conclude that the strategy of teacher identification is considerably better than random.

In French and physics Table 3.21 shows highly significant differences in all but one band in each subject. Although there was cause for concern in both of these subjects the Teacher and Test numbers were closer than the Test and Random allocation numbers. Again we can conclude that, even though there is cause for concern, a teacher strategy is considerably better than random.

Finally, in those few cells where there was little difference between the numbers chosen by teachers and by random allocation, we can reason that a strategy based on teacher's knowledge of some of a pupil's characteristics would be more likely to produce appropriate children than random allocation which took no account of these characteristics, so that the numbers selected may agree, but the quality of children chosen would not.

3.8 Subsets of the top 10% (top 5%, 2% etc.)

Firstly we can present Tables 3.22 and 3.23 which show the numbers of test identified (i.e. from rank ordering of the pupils based on predicted score, with no account of test error) pupils in the top 5% and top 2% who were also nominated by teachers, either in their top 5% or top 10%.

TABLE 3.22

PERCENTAGE OF TEST-IDENTIFIED TOP 5% PUPILS ALSO NOMINATED BY TEACHERS

	Percentage of test-identified top 5% nominated by teachers as coming into the	
	Top 5% group	Top 10% group
ENGLISH	53	71
FRENCH	36	57
PHYSICS	34	50
MATHS	47	69

TABLE 3.23

PERCENTAGE OF TEST-IDENTIFIED TOP 2% ALSO NOMINATED BY TEACHERS

	Percentage of test-identified top 2% nominated by teachers as:	
	Top 5% pupils	Top 10% pupils
ENGLISH	62	83
FRENCH	46	72
PHYSICS	48	67
MATHS	69	77

Both tables confirm the situation that has gradually emerged from previous sections, that French and physics teachers showed considerably less agreement with the tests than English and mathematics teachers, but in these results there is a large enough discrepancy in all subjects to cause some concern. Clearly, to be reasonably certain of including these smaller subsets of pupils in a teachers list of the more able, it is necessary to ask the teacher to identify a much wider group than just the 2% or 5% in which we might be interested. In addition to these conclusions we can add the following: when teachers completed their 10% lists they had the option to indicate any pupils whom they considered to be exceptional. Teachers varied enormously in their willingness to take up this option. It could be misleading to include, in the tables, figures that were based only on the 4 selected sample who took up the option. There was no clear evidence, however, that pupils who obtained the highest test scores in the sample as a whole were consistently classified as exceptional by their teachers. Thus, if we wished to ensure that such high scoring pupils were likely to be included amongst teachers nominations, it would seem to be inappropriate to ask the teachers to identify just an exceptional group. The unwillingness of teachers to make these final judgements of examiners ability further reflected the uncertainties that they had about this small sub-group of pupils. This has obvious implications for those who rely on teacher selection as a means of identifying small groups of outstanding pupils for activities such as summer schools. Recommendations relating to these inplications are made in Chapter Thirteen.

We would, of course, expect some mismatch between test and teacher because of test error, and we are able to make some comments with regard to this.

Firstly results that we have already discussed and those left for future discussion, that concern the A Group are comments about, roughly, the top 3 to 4%. Table 3.24 shows, over all schools, the percentage of pupils that fall into the A group.

TABLE 3.24

NOS. OF PUPILS IN THE A GROUP. TOTALS FOR ALL SCHOOLS

	ENGLISH	FRENCH	PHYSICS	MATHS
No. of pupils	79	88	69	63
Percentage of all pupils	3.39%	3.78%	2.96%	2.70%

Secondly, with regard to Table 3.17, the total number of pupils whose predicted score lies more than one standard error above the 90th percentile, represents about the top 23% of the pupils. Table 3.25 gives the figures for each subject.

TABLE 3.25

NOS. OF PUPILS WHOSE PREDICTED SCORES ARE MORE THAN ONE STANDARD ERROR ABOVE THE 90TH PERCENTILE

	ENGLISH	FRENCH	PHYSICS	MATHS
No. of pupils	45	69	35	36
Percentage of all pupils	1.93%	2.96%	1.5%	1.55%

We recall that from Table 3.19 we concluded that there were significantly more mismatches evident in physics and French than test error would allow. Thus, in these two subjects, we can again emphasise that teachers appear to be making mistakes in the pupils that they nominate for the top 10% even to the extent that they are missing a significant number in the top 5% and 2%.

Having discussed in some detail the relationship between test nominations and teacher nominations, and gained thereby some insights into the accuracy with which teachers can identify the top 10% of pupils in their own subject, we move on to the question of what it is that distinguishes the pupils who were picked by the teachers from those who were overlooked. This is the subject of several chapters of this report. The next chapter sets out what can be discovered by a study of the results of the two groups of pupils on the individual background tests. Later chapters tackle the issue in less statistical ways.

Notes for Chapter Three

1) A 'mismatch' would be a pupil who scored high on the test but was not identified by the teacher, or a pupil who scored low on the test but was identified by the teacher.

2) The standard error means that if the tests were applied to a large number of pupils and a subset of these pupils achieved the same test score, then the actual scores that the subset would achieve at O-level would be distributed around the score that they achieved on the test, with a standard deviation equal to the standard error. All that the test allows us to say with certainty about any one individual within the subset is that his actual O-level score will lie within this distribution. To be certain of where within this distribution his score will lie, we would need further information about the individual.

3) If two pupils X and Y had the same predicted score, there would be a 0.50 chance that the actual O-level score of X would be larger than that of Y. Thus if X has a lower predicted score than Y there will always be less than a 0.50 chance that his actual O-level score will be higher than that of Y. Thus a probability of 0.17 represents quite a high chance that reversal will take place.

4) The numbers in the Chi-squared analysis become rather small, and in conventional terms would render the analysis invalid. However Hopkins and Glass (1978, p322) suggest that the usual constraints on the test are too strict and in their terms the analysis remains sound.

5) We remind the reader, here that the 10% line and the other groupings were relative to the school norms and not to some wider local or national norms.

6) One row Chi-squared tests have been used (Minium, 1970). For example, one test, on the pupils in the band of scores from 1-1½ standard deviations below the 90th percentile in French, would compare the numbers of 20 pupils chosen by teachers and 140 pupils who were not chosen by teachers, with the test estimate of 13 pupils who could be expected to be chosen and 147 pupils who could be expected not to be chosen. Figures are taken from Tables 3.17 and 3.18. The test numbers are taken as the 'expected' values with which the teachers' numbers of

mismatches are compared in each of the bands. Cells are combined where necessary to give large enough numbers for a valid test.

CHAPTER FOUR

A Preliminary Investigation of the Characteristics of Children who were Nominated as More Able by their Teachers

4.1 Introduction

The analyses that are reported in this chapter are complementary to those reported in Chapter Three. They were carried out on the data from the same round of identification by teachers: that is the round in which all of the teachers involved in the research were given the task of nominating the top 10% in their subject without the help of checklists.

In Chapter Three we addressed the question of the extent of the agreement between teacher-based and test-based judgements of the potential of pupils for O-level performance in each of our four subjects. We were also able to obtain insights into the accuracy of teacher-based identification. In this chapter we attempt to explore a different question, namely: what is it that sets the pupils who were chosen by teachers apart from those who were not? This question was tackled in two ways. First, what sets teacher-selected pupils apart from the other pupils in the whole of the rest of the year group? Secondly, what sets teacher-selected pupils apart from those other pupils who were not identified by the teachers but who, nevertheless, had similar high potential in the subject according to the tests? Both aspects of the question were tackled by comparing the scores, on each background test, of pupils who were selected by teachers, with the scores of appropriate pupils who were not. In this way, distinguishing characteristics of the teacher-selected pupils could be established in each school, and patterns could be looked for across all schools. Such patterns could be expected to reflect any consistent bias which was colouring teacher judgement.

In the course of our discussions in schools, many teachers expressed concern that their overall judgement of pupils may have been unduly influenced by their own rough stereotypes of a more able child. Thus, for example, in identifying pupils as members of the top 10% group, a teacher might be biased towards pupils who did neat work, because these pupils happened to conform to his/her own stereotype of how a more able child should work. An interesting feature of this concern was that although teachers were aware that they might be referring to some rough stereotype they were not confident of their ability to explain the detailed nature of that stereotype.

It would seem, from other studies (Argyle and McHenry, 1971; Locksley et al, 1980) that crude stereotypes are likely to lead to a noticeable bias in the actual judgements made by a person in those situations in which that person lacks other, more relevant information about the individuals being assessed. We argue later (especially in Chapter Thirteen) that teachers do suffer from a shortage of relevent information in carrying out the particular task with which we are concerned in this report. We suggest in the later chapter that the nature of third year lessons (and of the written work which is based upon them) is such that able pupils have little opportunity to demonstrate to their teacher the depth and range of their abilities in the subject. It would therefore seem that the components of a teacher's stereotype of a more able child (e.g. that neat children are able) are quite likely to be expressed in that teacher's actual judgements under the particular circumstances in which the judgements about high ability are made. This would imply that the concern that teachers expressed about the influence of stereotypes is indeed one to be taken seriously.

An important aspect of the influence of stereotypes is that they may lead to associated biases in teachers' judgements in an entirely unconscious way. It should not be imagined that, in terms of our example above, the teacher notices that a child produces neat work and then makes a conscious decision that this neatness implies that the pupil is of high ability. It is more likely that the neat work is noticed by the teachers simply as 'good' work, and that the child is judged to be of high ability without the crucial factor of neatness ever coming to the teacher's conscious attention. Cook (1979, p23) provides further examples, and some discussion of this unconscious influence of stereotypes on overall judgements.

The possibility that the stereotypes may well influence teachers' judgements, and that this influence may well be unconscious implies that any insights that can be gained into the nature of the stereotypes, or of the biases that they produce, will be valuable for teachers would, on their own, find it difficult to be aware of the detailed nature of these effects. Without such awareness it is of course very difficult to begin to correct for the effects of the stereotype. It is this line of argument which provides the rationale for the work of this chapter.

4.2 Summary of conclusions

Before giving a detailed discussion of this work we will present a brief summary of the main results.

Comparisons between the selected pupils and all other pupils in the year group.

If the teachers' selection of pupils in a given subject is reasonably accurate, we would expect the selected pupils to have higher scores than the rest of the year group on many of the ability scales of our test battery, and, particulatly, to have higher scores on those scales which are appropriate to the measurement of potential in that subject. Where no difference could be found between the scores of the selected and non-selected pupils on a given variable we could draw a fairly strong conclusion that this variable did not influence teachers' judgements. This follows from the fact that the range of scores of the non-selected pupils in the whole of the year group will be large. If the selected pupils show no difference from the non-selected pupils, the implication is that the selected group also contains something approaching the full range of scores on that variable - clearly a result which would be inconsistent with a view that this particular variable greatly influenced teacher judgement.

The main result from this type of comparison were:

a) as expected, the pupils selected for the top 10% group scored higher on a large number of test scales. Higher scores on the majority of DAT scales were discovered for the pupils selected by teachers in all schools, This was the case in all subjects. Differences occured in particular schools and in particular subjects on the other variables (Creativity, Neatness, Social Class, Attitudes and Sex). (See Table 4.2)

b) The strongest biases that could be detected by the test scales, varied from school to school within each subject. In the case of a small number of teachers, some unexpected and apparently inappropriate characteristics emerged. However in general the characteristics of the selected groups were not very surprising. What might be thought of as favourable pupils attitudes, (particularly the possession of a good Academic Self Image in mathematics and physics, and in all subjects, though to a lesser extent, the possession of a positive Relationship with the Teacher), were characteristic of a considerable number of teacher-selected groups. Most striking, however, was the influence of the general ability variable VN in all subjects. This was a particularly strong characteristic of the teacher identified group in physics. This, rather than the other DAT scales (such as Mechanical Reasoning in physics) that appeared to be more appropriate as indicators of potential in the individual subjects, dominated as the characteristic of the teacher identified group. (See Table 4.4)

c) Only in very few cases were there differences between the selected and the non-selected pupils on the variables of social class and neatness that could not be accounted for by links between these variables and other variables. (See Table 4.4). In view of the argument that this piece of work is especially useful in indicating variables that have little effect on teachers' judgements, we would suggest that our evidence tends to contradict the view that teachers in individual schools were strongly influenced by these factors.

d) Only in French was there a moderate number of individual schools in which there was a sex bias (in favour of girls) in the groups selected by teachers. (See Table 4.2) Again there is an implication that teachers do not show a strong bias for any one sex in most individual schools, in most subjects.

e) There was a variation from subject to subject in the extent to which pupils opinions of the subject differed, between the teacher-identified pupils and the rest of the year group. In mathematics and English, opinions as to usefulness of the subject did not show marked differences, whereas in French and physics they did. In every subject a higher proportion of the pupils chosen by their teachers tended to think that they were good at the subject; this was most strongly the case in French and mathematics. Liking the subject also strongly characterised the pupils chosen in French. This was a weaker characteristic of the chosen pupils in mathematics and physics, and was a noticable characteristic in only a small number of schools in English. (See Table 4.6)

f) Very few meaningful differences were detectable between schools of the various types. This enhances the conclusion that different teachers showed different individual biases, and that this variation was largely independent of the broad organisational charcteristics that we used to group schools into different types.

Comparisons between selected pupils and other pupils from the B Group who were not selected.

If teachers' identification strategies were sound we would expect the pupils selected in a given subject to differ from the non-selected B Group pupils on ability test scales, and in particular on the test scales that were appropriate to the measurement of potential in that subject. However, within the B Group, the range of scores on the most appropriate test scales was relatively small (because the B Group was designed specifically as a group with a fairly similar measured potential for the subject). It was therefore quite difficult for teachers to distinguish between the pupils with the highest measured potential and the others in the group with slightly lower measured potential. The range of scores on other scales - which did not relate to the overall measure of potential - was however rather greater. We might therefore expect that, if a characteristic measured by one of these other scales did tend to influence a teacher's judgement, he might be particularly inclined to respond to the considerable variation in this characteristic in selecting pupils from the B Group. We might suppose therefore that an analysis of selection from the B Group might be especially fruitful in bringing to light the nature of any more or less irrelevant characteristics that affected teachers' judgements. Thus this analysis is complementary to that involving comparison with the whole year group which, as we have argued, was particularly effective at bringing to light characteristics which did not have significant influence on teachers' judgements.

The findings from the B Group analysis were:

a) We found fewer significant differences on background scales between teacher-selected pupils and the rest of the B Group than were found when the pupils identified by teachers were compared with the pupils in the whole of the rest of the year group. This suggested that the factors that did affect the way in which teachers made their decision about which B Group pupils to include on their top 10% lists, were factors that had not been measured by our test battery.

b) The characteristics that were identified varied from school to school within each subject.

c) In physics, in most schools, higher scores on the general ability scale (VN) featured as a characteristic of the groups of pupils indentified as more able by teachers.

d) In French, there were strong indications that pupils were chosen who either liked the subject or thought they were good at it. These indications were present but not so strongly in English, but were not present in mathematics or physics.

e) Rather as expected, many of the factors which distinguished the teacher-identified children from the other members of the B Group did not seem to be particularly appropriate. There did seem therefore to be some evidence that inappropriate biases were entering into the teachers' identification strategies.

f) There was some evidence that the schools in which the teachers appeared to have a good selection strategy (in that the characteristics of the chosen pupils did seem to be appropriate) were the 13+ entry schools and also the smaller schools.

Two further kinds of analysis were carried out:

1. When comparisons were made within the 'A' Group[1]
   a) Very few significant differences between mean scores of those chosen, compared with those not chosen, were discernable in any subject. Only the pupils 'Attitudes to the Subject' characterised the pupils who were chosen. This was particularly true in French.

2. When comparisons were made in the C group:
   a) Significant differences associated with general ability (VN) were evident in physics, French and mathematics, between the pupils who were chosen and those who were not. Differences in attitudes were also detectable in those subjects. In French and physics strength in the variables SR (Space Relations), SP (Spelling) and LU (Language Usage) further characterised the pupils who were chosen by their teachers.

   b) By comparing the numbers of boys and girls chosen with the numbers who were not it was clear that significantly more C Group girls than boys were chosen in French. There was no significant bias in the other subjects.

The indications in all aspects of this work were that individual bias of judgement was present from teacher to teacher. Further bias could exist in terms of variables which did not form part of our background test battery so that, though the work reported here gave important clues to the nature of teacher-based identification, a further study of a different kind would be necessary in order to gain fuller understanding. Two appproaches to this issue of further study are discussed in later chapters of this report.

Although differences between the chosen and not chosen pupils, on such variables as sex and social class, were rarely strong enough to reach the level required for statistical significance in any one school, certain trends were evident in the data. These trends tended to be consistent across the whole sample of schools and they were shown to give rise to some significant differences between the chosen and the not chosen pupils when the results of all of the schools were pooled. In all subjects a cause for concern could be identified as a result of some of these comparisons: a sex bias (towards girls) was seen; a social class bias was shown to be present; and biases with regard to attitudes and neatness were established. These biases were by no means as dominant as is sometimes suggested in the literature or they would have emerged more strongly in the work done in each of the individual schools. However they should not be neglected, particularly in French where the strongest indications of this kind were found.

We will now give fuller account of our methods of analysis. The reader will see that there are many tables of figures that could be presented in connection with this part of our study. A number of these tables have been put into Appendix 4 to lighten the load somewhat. Appendix 1 should be consulted in order to interpret the test scale codes that we have used, where fold out page has been provided for the convenience of the reader.

4.3 Characteristics of the teacher selected group compared with the whole of the rest of the year group

Correlation analysis

This particular analysis is entirely *independent* of our test identification procedure which is based on multiple regression analysis because it refers to the individual background test scales and not, in any way, to the combinations of scales that we used as predictors of future O-level performance. It is therefore possible to interpret it as giving largely independent insights into aspects of the mechanisms that lie behind the results reported elsewhere.

The most productive technique used for this analysis was that of correlation (and, later, partial correlation) and the results of this kind of analysis therefore follow. To carry out this procedure we gave a score of +1 to those pupils who were selected by teachers and a score of 0 to those who were not. Using the SPSS package of computer programmes the Pearson Correlation Coefficients were calculated between this dichotomous teacher selection scale and each of the background test scales. (Under these conditions the Pearson coefficient is equal to the Point Biserial Correlation Coefficient which is the appropriate statistic for relating dichotomous to continuous scales.) (Guilford and Fruchter, 1978.)

Where a statistically significant positive correlation was established between teachers' judgement and one particular test scale (spelling, say) we could say that pupils in the teacher selected group generally had a higher spelling score. We could go on to infer that either good spelling itself, or something with which spelling was highly correlated was one of the things that may have been important in the overall identification strategy in that school.

Table 4.1 shows which variables were significantly correlated with teachers' judgement. (All such correlations were positive except in the case of social class where all were negative. Since social class 1 is the professional end of the scale and social class 5 the manual end, this negative correlation shows that teachers tended to select offspring of professional parents). All pupils in the sample were included in this analysis which was carried out separately for each subject in each school. While presenting this table as a well verified map of the characteristics of the teacher-selected pupils and as an indication (no stronger) of the overall selection strategy, we must be careful not to infer that it represents the detailed procedure which teachers followed.

In terms of our example above, the selected group were characterised by better spelling than the rest of the year group, and we can therefore infer that spelling, or a correlate of spelling, may have been a component in the overall selection procedure. However, we cannot imply that teachers actively sought out the good spellers. Accurate spelling may have unconsciously influenced the teachers, or it may have been relevant at an earlier stage in the selection process. This final point deserves some amplification.

The characteristics shown in Table 4.1 were characteristics on which the selected group were better than the whole of the rest of the year group. Teachers were however unlikely to be making their selection from the whole year group. They were much more likely to be selecting from a smaller sample - e.g. the top set. It may be that spelling influenced their selection at the stage when the top set itself was being chosen (rather than the stage at which the finer judgement about the top 10% was being made). If, for example, the setting procedure placed only good spellers in the top set, the teachers would be bound to construct a top 10% group characterised by high spelling scores irrespective of any bias for or against good spellers that they may show when they go about the task of nominating pupils for the 10% group.

However, it must be born in mind that the setting procedure is itself dependant upon teachers' opinions and will reflect their biases. It will be directly dependant on these biases if the setting is based on teachers' subjective opinion of pupils. It will still be indirectly dependant upon their views even if it is based on a test or on pupils' classwork marks. In these cases teachers biases will colour the choice of the test (e.g. will influence the decision as to whether a general IQ test or a subject specific set of tests should be used) and will have a considerable influence on the kinds of criteria which are used in the assessment of classwork. Despite the complication of the role played by setting procedures in the process of identifying the top 10%, we suggest that the characteristics of the top 10% groups that we set out in the following pages do give insights into the nature of the biases which affect teachers' judgements. The point that we are making here is that we cannot distinguish with certainty between biases that have their effect at the stage when the top sets are produced, and biases that have their effect when teachers are engaged on the finer selection of the 10% from the top set.

In Chapter Six, where we report biases apparent in the second round of identification, we can take this discussion forward in the light of the new information and attempt to make some distinction between the two points at which bias can affect the process of identification.

The characteristics reported in Table 4.1 therefore reflect this overall process of selection in which the placing of a name on the top 10% list was only the final step.

TABLE 4.3 Test Scales which showed significant correlation with teacher selection

		CREATIVITY			ATTITUDES								DIFFERENTIAL APTITUDE TESTS											
	SCHOOL	FLU	FLEX	ORIG	A	B	C	D	E	F	G	H	VR	NA	AR	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	SC
ENGLISH	00				*	*	*				*		*	*	*	*	*	*	*	*	*			*
	01	*	*	*	*	*	*	*	*			*	*	*	*	*	*	*	*	*	*	*		*
	02			*	*		*	*			*	*	*	*	*	*		*		*	*			*
	03	*	*	*		*	*	*			*		*	*	*	*	*		*	*	*	*		*
	04		*		*	*	*		*		*	*	*	*	*	*	*	*	*	*	*	*		
	05	*	*	*		*	*		*				*	*	*	*		*	*	*	*	*		*
	06	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*			
	07		*	*	*	*	*		*	*	*	*	*	*	*	*	*	*	*	*	*		*	
	08		*			*	*	*			*		*	*	*	*	*	*	*	*	*			
	09	*	*	*	*	*	*		*		*	*	*	*	*	*	*	*	*	*	*			*
	10		*		*	*	*				*		*	*	*	*	*	*	*	*	*	*		*
	11					*	*		*	*			*	*	*	*	*		*	*	*	*		
	12		*			*	*						*	*	*	*	*	*	*	*	*			
FRENCH	00					*	*				*		*	*	*	*		*	*	*	*			*
	01	*	*	*	*	*	*	*	*		*	*	*	*	*	*	*	*	*	*	*	*		*
	02			*	*	*	*	*	*		*	*	*	*	*	*		*		*	*			*
	03	*	*		*	*	*				*	*	*	*	*	*	*		*	*	*	*	*	
	04		*		*	*	*	*	*		*	*	*	*	*	*		*	*	*	*	*	*	
	05	*	*	*			*						*	*	*	*	*	*	*	*	*	*		*
	06	*	*	*	*	*	*	*		*	*	*	*	*	*	*	*	*	*	*	*	*	*	
	07		*	*	*	*	*				*	*	*	*	*	*	*	*	*	*	*		*	*
	08		*	*		*	*	*	*		*	*	*	*	*	*	*		*	*	*		*	
	09	*	*	*	*	*	*		*		*	*	*	*	*	*	*	*	*	*	*			*
	10		*			*	*	*			*		*	*	*	*	*	*	*	*	*			*
	11					*	*		*	*	*	*	*	*	*	*	*		*	*	*	*		
	12			*				*					*	*	*	*	*	*	*	*	*			
PHYSICS	00				*	*	*	*			*		*	*	*	*		*	*	*	*			*
	01			*	*	*	*		*				*	*	*	*		*	*	*	*			*
	02	*	*	*	*	*	*	*	*			*	*	*	*	*	*	*	*	*	*			*
	03		*	*			*						*	*	*	*		*	*	*	*			
	04		*		*	*	*		*				*	*	*	*	*	*	*	*	*	*		
	05	*	*	*	*	*	*			*		*	*	*	*	*		*	*	*	*			*
	06	*	*	*	*	*	*	*		*	*	*	*	*	*	*	*	*	*	*	*			*
	07			*	*	*	*			*	*	*	*	*	*	*		*	*	*	*			*
	08	*	*	*		*	*	*			*	*	*	*	*	*	*	*	*	*	*			
	09	*	*	*	*	*	*		*		*	*	*	*	*	*	*	*	*	*	*			
	10				*	*	*	*			*		*	*	*	*		*	*	*	*			*
	11			*			*			*			*	*	*	*	*	*	*	*	*			
	12		*				*						*	*	*	*		*	*	*	*			*
MATHEMATICS	00	*	*	*			*						*	*	*	*	*	*	*	*	*			*
	01	*	*	*	*	*	*	*			*		*	*	*	*	*	*	*	*	*			*
	02				*	*	*	*			*	*	*	*	*	*	*	*	*	*	*			*
	03		*		*	*	*				*		*	*	*	*	*	*	*	*	*	*		*
	04	*	*	*	*	*	*	*	*				*	*	*	*	*	*	*	*	*	*		
	05		*			*	*					*	*	*	*	*		*	*	*	*			*
	06			*	*	*	*	*			*	*	*	*	*	*	*	*	*	*	*			
	07		*	*	*	*	*				*	*	*	*	*	*	*	*	*	*	*			*
	08		*	*		*	*	*	*		*		*	*	*	*	*	*	*	*	*			
	09	*	*	*	*	*	*		*			*	*	*	*	*	*	*	*	*	*			*
	10					*	*				*		*	*	*	*	*	*	*	*	*			
	11						*			*			*	*	*	*	*	*	*	*	*			
	12		*	*	*	*	*		*	*			*	*	*	*	*	*	*	*	*			*

NOTES: 1) The chi-squared statistic was used to determine significant correlations between teacher selection and sex, neatness and social class. The significance of the correlation coefficient was used for all other scales. Variables are ticked if the level of significance is 5% or higher (i.e. 1%, 0.1% etc.).

2) Correlations for sex always indicated that teachers tended to select girls.

It is not surprising to find that, measured against the whole of the year group, there were significant correlations between many of the individual test scales and teacher selection. To help the reader to a clearer understanding of the implications of Table 4.1 we have summarised the data, to show, for each background scale, in how many schools a significant correlation had been revealed by the analysis. This summary is presented in Table 4.2.

Table 4.2 gives an indication that there was some variation from school to school in the overall characteristics of the selected groups. There was, however, a strong tendency for teacher selection to be correlated with scores on the cognitive test scales (the Differential Aptitude Test Scales). Only Attitudes B and C ('Relationship with Teacher' and 'Pupils' Academic Self Image') were consistently correlated with teacher selection. Schools differed in the extent to which the other attitudes and the three creativity scales were correlated with teacher selection. It is perhaps interesting that high social class was a factor in some subjects in some schools (just over half of all possible areas of selection showed significant association between teacher selection and this variable). However, social class was not a consistent correlate of selection. Neither was sex. On this analysis sex differences showed up mainly in French where teachers in 5 schools tended to pick girls.

TABLE 4.2

NO. OF SCHOOLS* IN WHICH THERE WAS SIGNIFICANT CORRELATION BETWEEN THE SELECTION OF PUPILS AS 'MORE ABLE' BY THE TEACHER AND THE PUPILS' SCORES ON EACH BACKGROUND TEST

VARIABLE	ENGLISH	FRENCH	PHYSICS	MATHS
CREATIVITY				
FLU	5	5	5	4
FLEX	10	9	8	9
ORIG	7	8	9	8
ATTITUDES				
A	8	8	9	8
B	12	11	10	11
C	13	12	13	13
D	5	7	5	5
E	7	6	4	4
F	3	2	4	2
G	9	11	6	7
I	6	9	6	5
DAT				
VR	13	13	13	13
NA	13	13	13	13
VN	13	13	13	13
AR	13	13	13	13
CSA	11	10	6	12
MR	11	10	13	13
SR	12	12	13	13
SP	13	13	13	13
LU	13	13	13	13
SEX**	1	5	0	0
NEATNESS (NT)	6	6	1	2
SOCIAL CLASS (SC)***	7	7	8	8

*Each section of school 13 was analysed separately. The maximum number of schools which could show significant correlations on any variable was therefore 13.

**Only 11 schools were mixed. The total number of schools analysed with respect to sex was therefore 11. All correlations with sex were such as to indicate that teacher selected groups were biased towards girls.

***All correlations with social class were negative indicating a teacher group bias towards the professional end of the social spectrum.

Partial correlation analysis

Important though they are, the results presented above are only a starting point for understanding the correlates of teachers' selection. The picture is complicated by the fact that pupils' scores on all of the individual test scales will be correlated with one another, so that a correlation between teacher selection and one variable will automatically generate a correlation (probably a weaker correlation) between teacher selection and another variable.

It is possible to scrutinise the correlation matrices and pick out the variables that show, for each subject, in each school, the strongest links with teacher selection, but a better approach is that of partial correlation analysis. This is the process whereby an adjustment of scores on all but one of the background tests is made to compensate for differences in score on the one remaining test variable (the control variable). The degree of correlation between the adjusted scores and teacher selection then indicates the strength of these links once any effect due to the control variable has been removed.

In this way, those variables which consistently show significant correlations with teacher selection even when scores are adjusted, in turn, for differences in all other variables can be brought to light[2]. These variables indicate the strongest characteristics of the teacher selected groups.

In Table 4.3 we indicate the nature of these strong characteristics. Comparison with Table 4.1 shows that many characteristics have 'dropped out'. These can be regarded as the characteristics which were revealed by correlations generated automatically by the inter-relationship between the background variables.

TABLE 4.3

MAIN VARIABLES CHARACTERISING SELECTED GROUP SHOWN BY PARTIAL CORRELATION

SCHOOL	ENGLISH	FRENCH	PHYSICS	MATHS
00	(VR and NA) or SR or LU	VN	C or VN or SR	NA or SR
01	VN or LU or ORIG	(VN and NA)	(NA and VN) or MR	VN
02	SP or LU	(LU and SP or VN or (C and D)	VN or C or E	NA or C
03	FLEX or VN or SP or B	LU or NA or or CSA	VN or MR	NA or C or SC
04	LU or SP or B	LU or A or B or NT	SR or A or LU	NA or C
05	VN or C or SP	VN or SC or (SP and LU)	VN or FLEX or C or MR	NA or C
06	LU or B or C or FLEX or NA	LU or B	VN or B or C	NA or C
07	VN or C or SP or LU	VN or LU or SP or A or C	VN or SR	VN or C or SR
08	VN or SP	LU and VN	VN or (FLEX and ORIG) or B	NA
09	VN or LU or ORIG	ORIG or UR or (SP and LU)	VN or LU or MR or C	VN or LU
10	VN or LU or SC	LU or SC	VN or B or C	VN or SC
11	VN or B	VN or NT	VN or SR	VN or NA or VR or SR
12	LU or SP or NA	SP	VN	VN

The use of statistics of this kind draws us away from the decision process that the teachers used when committing names to a 'top 10%' list. We would be wrong to think that they went through a process like partial correlation. Indeed, as we remarked before, they would probably not even have been aware of making an overall selection from the whole year group. Nevertheless, the characteristics of the group brought to light in this way should give clues to the underlying strategy and we may be able to comment on the final result of the selection process irrespective of the route taken to achieve it. The Pearson Correlation Coefficients for the variables in Table 4.3 are given in Apppendix 4.

It is not surprising that the strong characteristics of the group aried from subject to subject (this matched our overall subject specific viewpoint which was brought out by the regression equations, in that they were different for each subject). It is surprising that the characteristics differed so much from school to school and that they were not always the expected indicators of ability (e.g. we found NA to be a characteristic for French in School 01, LU for mathematics in school 09). On the basis of the regression equations one would expect in every school the table to show LU as a strong characteristic in English and French, MR and NA in physics, NA and possibly SR in mathematics. However, good attitudes were strong characteristics in some schools, and some subjects. Neatness, sex, creativity and social class were also strong characteristics here and there (though they were by no means dominant).

None of the sets of characteristics stand out as completely 'wrong' and certainly a closer study of individual pupils, in terms of their test scores, and of their work in the classroom would be needed before one could draw final conclusions, but one is led to the tentative view that the identification, even at this overall level of 'top 10%' from whole year group needs some 'finer tuning' in all subjects. The biases of different schools in each subject would indicate that this 'fine tune' is different for each identification process.

It is instructive to draw up a table like Table 4.2 to summarise th e results found for the whole sample of schools.

TABLE 4.4

NO. OF SCHOOLS SHOWING EACH TEST AS A STRONG CHARACTERISTIC OF ITS 'TOP 10%' GROUP AFTER PARTIAL CORRELATION

VARIABLE	ENGLISH	FRENCH	PHYSICS	MATHS
CREATIVITY				
FLU	0	0	0	0
FLEX	2	0	2	0
ORIG	2	1	1	0
ATTITUDES				
A	0	2	1	0
B	4	2	3	1
C	3	2	6	5
D	0	1	0	0
E	0	0	1	0
F	0	0	0	0
G	0	0	0	0
I	0	0	0	0
DAT				
VR	1	1	0	1
NA	2	2	1	8
VN	9	7	12	6
AR	0	0	0	0
CSA	0	1	0	0
MR	0	0	4	0
SR	1	0	4	3
SP	7	4	0	0
LU	9	9	2	0
NEATNESS (NT)	0	2	0	0
SOCIAL CLASS (SC)	1	2	0	2

Sex, being a categorical variable, is not really suited to this analysis and is omitted from this table, and from Table 4.3.

The general academic ability scale VN (VR and NA) was a characteristic in most subjects and in all schools for those identified. Physics in particular had this characteristic very strongly and, one might suppose, too strongly. The absence of MR in physics in all but four schools would indicate a bias towards a less valid criterion (VN) and away from the more valid one (MR). This may well be due to lack of real knowledge of pupils at third year level because of the limited teacher/pupil contact which is so often a characteristic of physics teaching in the early years of secondary schooling. The bias towards general ability may also be a result of the fact that work in science up to third form level may not be sufficiently challenging to pupils' specific scientific aptitudes, so that teachers do not get very much information about their pupils strengths and weaknesses in these areas.

In general one could conclude that in all subjects the presence of the VN components reflects an inability on the part of the teachers to pick up the more appropriate ingredients of ability in a confident manner. High attitudes, on one or other of the scales, were characteristic of nearly half of the identifications; high scores on Attitudes B (Relationship with Teacher) and C (Academic Self Image) occurring most frequently. High creativity as measured by Fluency, Flexibility or Originality did not characterise many of the groups.

Social class and neatness were not usually characteristics of teacher-selected pupils, once the effect of the correlations between these and other variables had been accounted for. Sex was a variable that could not reasonably be analysed in this way but its independant influence can be assumed to be no stronger than was indicated by the figures in Table 4.2. Only in French did a substantial number of the teachers in individual schools show a sex bias - and this was consistently in favour of the selection of girls. Since, as we have argued in the introduction to this chapter, this part of the work gives fairly strong indications of the variables which had little influence on teachers, we are led to conclude that the influence of social class, neatness and sex (variables which are often thought to be powerful in influencing teachers' judgements) was weaker than is commonly supposed: and in fact that it was rarely strong enough to produce noticeable biases within the individual schools.

On the variables included in Tables 4.2, 4.3 and 4.4, no differences that relate to types of school were apparent. For example, the 13+ entry schools (02,05,09) did not show markedly different characteristics from the rest of the sample.

We were able, via a questionnaire to pupils, to obtain some information on their likes and dislikes, on whether they thought a subject was useful or not, and on whether they thought they were good at a subject[3]. These factors could have influences on their performance in a subject, and could affect the teachers' judgement of their ability. In seeking to bring such influences to light, we carried out chi-squared tests on contingency tables similar to those which are given, as examples, in

Table 4.5[4]. For each subject, we counted how many chosen, and how many not-chosen pupils gave a positive response, by including the subject in their answer to a request such as:

'List the subjects which you think you are good at'.

We also counted how many gave a neutral response by not including the subject. A statistical comparison was then made to establish if there was a difference between those pupils appearing on the teacher's list and those not.

TABLE 4.5

RESULTS FOR SCHOOL 04 ENGLISH

LIKING OF SUBJECT	No. who did not say they liked subject	No. who like subject	Expected Table		
TEACHER IDENTIFIED	2	16	7.6	10.4	CHI-SQUARED = 6.50
NOT TEACHER IDENTIFIED	72	86	66.4	91.4	PROB. .05(*)
DISLIKING OF SUBJECT	No. who did not say they disliked subject	No. who disliked subject	Expected Table		
TEACHER IDENTIFIED	18	0	12.5	2.5	FISHER EXACT PROB. = .006(**)
NOT TEACHER IDENTIFIED	107	24	109.5	21.5	
THOUGHT SUBJECT USEFUL	No. who did not say they thought it useful	No. who thought it useful	Expected Table		
TEACHER IDENTIFIED	2	16	5.1	12.9	CHI-SQUARED = 2.070
NOT TEACHER IDENTIFIED	48	110	44.9	113.1	NSD
THOUGHT SUBJECT NOT USEFUL	No. who did not say it was not useful	No. who thought subject not useful	Expected Table		
TEACHER IDENTIFIED	18	0	17.2	0.8	FISHER PROB. = 0.414
NOT TEACHER IDENTIFIED	150	8	150.8	7.2	NSD
THOUGHT GOOD AT SUBJECT	No. who did not say they were good at subject	No. who thought they were good at subject	Expected Table		
TEACHER IDENTIFIED	5	13	12.7	5.3	CHI-SQUARED = 15.334
NOT TEACHER IDENTIFIED	119	39	111.3	46.7	PROB. .001(***)

No differences according to type of school could be detected and the analysis by school and subject is summarised in Table 4.6 where we quote the number of schools (out of 13)[5] that showed significant difference[6] between the teacher selected group and the rest of the third year on each question.

TABLE 4.6

NUMBER OF SCHOOLS SHOWING SIGNIFICANT DIFFERENCE ON QUESTIONS RELATING TO SUBJECT POPULARITY, WHEN THE TEACHER SELECTED GROUP IS COMPARED WITH THE REST OF THE YEAR GROUP

	LIKE/DISLIKE	USEFUL/NOT USEFUL	THINK GOOD AT
ENGLISH	3	0	6
FRENCH	12	9	13
PHYSICS	8	7	8
MATHS	7	0	10

Differences were always in the expected direction for the questions, i.e. the teacher-selected group tended to like the subject, thought the subject useful and thought that they were good at the subject.

There is strong evidence that French selection was biased towards those who liked French and those who thought they were good at it in almost all schools. Pupils opinions of their ability in mathematics also seemed to be generally important.

The noughts in the USEFUL/NOT USEFUL column for English and mathematics perhaps reflected a tendency of both identified and not identified groups to think the subjects useful. It is instructive to reflect on the results in relation to a comparison made when the figures for individual schools were summed to indicate the presence of any trend in the results of the sample as a whole. These results are shown in Table 4.7.

TABLE 4.7

COMPARISON OF TEACHER IDENTIFIED PUPILS WITH REST OF YEAR GROUP TOTALS FOR ALL SCHOOLS

	SAID 'LIKE'	NOT	SAID 'DISLIKE'	NOT	SAID 'USEFUL'	NOT	SAID 'NOT USEFUL'	NOT	SAID 'GOOD AT'	NOT
ENGLISH										
TEACHER IDENTIFIED	172	48	10	210	189	31	3	217	144	76
NOT TEACHER IDENTIFIED	1202	736	423	1515	1515	423	47	1891	673	1265
FRENCH										
TEACHER IDENTIFIED	146	73	30	189	129	90	13	206	112	107
NOT TEACHER IDENTIFIED	1416	523	958	981	495	1444	475	1464	287	1652
PHYSICS										
TEACHER IDENTIFIED	132	87	41	178	88	131	22	197	90	129
NOT TEACHER IDENTIFIED	700	1239	724	1215	653	1286	296	1643	328	1611
MATHS										
TEACHER IDENTIFIED	164	59	26	197	207	16	0	223	154	69
NOT TEACHER IDENTIFIED	960	975	598	1337	1608	327	49	1886	632	1303

The comparisons that showed significant differences according to the Chi-squared test are shown in Table B20.

TABLE 4.8

SIGNIFICANT DIFFERENCES FROM TABLE 4.7 COMPARISONS

	'LIKE'	'DISLIKE'	'USEFUL'	'NOT USEFUL'	'GOOD AT'
ENGLISH	***	***	**	NSD	***
FRENCH	***	***	***	***	***
PHYSICS	***	***	NSD	***	***
MATHS	***	***	***	*	***

* = $p \leq .05$

** = $p \leq .01$

*** = $p \leq .001$

The overall trend, therefore, when the results from individual schools were combined, was that significantly more pupils amongst those who were identified by teachers as able, liked the subject, found it useful and thought that they are good at it, than was the case amongst those pupils in the rest of the year group who were not identified by teachers. This was true in all four subjects.

In all subjects, nevertheless, a large number of the teacher-selected pupils did not say that they thought themselves good at the subject. This was particularly so in French and physics.

One wonders how much this difference in pupils' views of their own ability actually affected their performance in the subject. It is also interesting to note the large numbers of those who were not identified who thought themselves good at the subject. This could of course be because their assessments were made relative to the expectations within their own particular class.

In conclusion, it would appear that there is not a great cause for concern about the characteristics of the teacher selected group compared with the whole school. There is, however, some concern that emphasis was not always strong on the most relevant characteristics. General academic ability was present more strongly than one would hope if judgement of ability was adequately sensitive to the special characteristics of high potential in each subject area. Individual teachers did seem to have individual biases and there was no one type of school environment in which teachers were able to be particularly successful in identifying the more able pupil. Social class, neatness and sex did not appear to have an overwhelming influence on teachers' judgements in the individual schools.

Distribution of abilities for the teacher selected group compared with the distribution for all other pupils

This work is complementary to that reported in the previous sections but provides the opportunity for an interesting development of the analysis which is reported in the following sections.

This work is based on the technique of dividing the scores on each of the background tests into three categories: score below 50th percentile, score between 50th and 84th percentile, score above 84th percentile (on local norms)[7]. For each school, and each subject, the numbers of pupils, from the teacher-selected group, whose scores fell into each of these categories were determined, and thus the distribution of the scores of the teacher-selected pupils across the three categories was established. This distribution was then compared, using the chi-squared test, with the distribution of the scores of the non-selected pupils across the same categories. This study provided confirmation of the findings reported in Table 4.2. This can be seen from Table 4.9 where we indicate, for each test scale in turn, the number of schools in which the high proportion of high scoring pupils in the teacher selected group, was sufficient to generate a significant difference between the test score distributions of the teacher-selected pupils and the rest of the year group.

The finding that teacher selection was consistently associated with high DAT scores and that, with the exception of Attitudes B and C, other associations occured in odd schools but not consistently across the sample, is even clearer here than it was in Table 4.2. This can be seen from the high numbers in the DAT section of the table and in the lines corresponding to Attitudes B and C (especially C). In the rest of the table the numbers are small indicating that, on these scales, only in an odd school here and there were the teacher-selected pupils significantly higher scoring than the rest.

TABLE 4.9

NO. OF SCHOOLS SHOWING A SIGNIFICANT DIFFERENCE IN DISTRIBUTION ON EACH VARIABLE BETWEEN THE TEACHER SELECTED AND THE NON SELECTED PUPILS

	No. of schools showing difference in distribution in:			
VARIABLE	ENGLISH	FRENCH	PHYSICS	MATHS
CREATIVITY				
FLU	3	1	0	1
FLEX	6	5	5	6
ORIG	7	4	5	2
ATTITUDES				
A	5	4	3	4
B	9	8	7	8
C	10	7	10	11
D	3	3	3	3
E	2	2	3	0
F	0	1	0	1
G	3	11	3	0
I	6	4	0	1
DAT				
VR	13	13	13	13
NA	13	12	13	13
VN	13	13	13	13
AR	13	13	12	13
CSA	7	5	4	7
MR	9	5	13	11
SR	10	8	13	13
SP	13	13	13	13
LU	13	12	13	13
SEX[**]	1	5	0	0
NEATNESS (NT)	6	6	1	2
SOCIAL CLASS (SC)[***]	7	7	8	8

---

[*] For these categorical variables the figures here were identical to those in Table 4.2, because for these variables, the technique used for Table 4.2 was essentially the same as that used here.

Differences between schools and groups of schools in the distribution of scores in the teacher selected groups

Though the confirmatory nature of the results of Table 4.9 were encouraging, the main interest in establishing the distribution of test scores within the teacher-selected groups in each school was that it provided a basis on which these groups could be compared across schools. By doing such a comparison, we determined whether the nature of the selected groups were significantly different (in terms of the test scales) from one school to another. The variety of characteristics associated with teacher selection in different schools (as revealed in Tables 4.1 and 4.3) suggested that there would be differences on some scales, but this alone could not indicate whether these differences were strong enough to be significant.

To test this possibility we therefore made, for each background test, in each subject area, three comparisons of the distributions of scores in the teacher-selected groups:

i) comparison between the schools individually;

ii) comparison between 11+ entry and 13+ entry schools;

iii) comparison between large, medium and small schools in the 11+ entry group

Individual school differences

Before presenting the results of this analysis an example of the procedure is set out in some detail.

Let us take the pupils who were selected by teachers in English and look at their scores on the Flexibility test. In school 00, on this test, 6 selected pupils had scores below the 50th percentile, 5 had scores between 50th and 84th percentile and 3 had scores above the 84th percentile. Thus the distribution of scores across these categories can be given by the numbers 6, 5 and 3.

Table 4.10 shows similar distributions for all schools.

TABLE 4.10

NUMBER OF PUPILS SELECTED BY TEACHERS AS IN TOP 10% IN ENGLISH WHOSE FLEXIBILITY SCORES FELL IN THE GIVEN RANGES

School	Score range in percentiles 0 - 50th	50th - 84th	84th - 100th
00	6	5	3
01	2	9	7
02	11	4	4
03	0	7	8
04	6	8	4
05	4	6	6
06	4	5	11
07	8	9	5
08	2	5	5
09	3	13	8
10	4	4	4
11	4	4	2
12	3	6	3

If this table is subjected to chi-squared analysis the significance of the difference in the distributions in the thirteen schools can be assessed.

For this table the chi-squared value was 32.5 which, for a table with 24 degree of freedom, failed to reach the 5% level of significance. We therefore conclude that on this test, teacher-selected pupils in English were distributed across the three score categories, in all schools, in comparable ways.

Clearly the numbers in Table 4.10 were fairly small and for some variables fell sufficiently low that the validity of the chi-squared analysis was brought into question. In such cases columns 1 and 2 or columns 2 and 3 were combined to restore the numbers to a more appropriate level. (Such combination reduced the number of degrees of freedom for the chi-squared analysis to 12. Note is made of this at appropriate points in the following table.)

In that table we set out the chi-squared values for all of the comparisons like the one described above (i.e. for all test scales and in all four subjects). The overall conclusion is that there were no significant differences between schools in the distribution of scores of the teacher-selected pupils on the Differential Aptitude Scales, nor were there differences on Attitudes B or C. On the other scales significant differences were sometimes apparent. This is in line with the finding that, on the DAT scales, and on Attitudes B and C all schools showed a correlation between teacher selection and the appropriate test score, whereas on the other scales only some schools showed a significant correlation.

It is interesting that school 07, a school with a fairly difficult catchment area had teacher selected groups in all subjects in which more pupils than usual for the sample as a whole had high scores on the 'social adjustment' scale (Attitude E). On this same scale school 06 had a lower proportion of high scorers in its teacher selected groups. School 06 has a more favourable catchment area[8].

TABLE 4.11

SUMMARY OF CHI SQUARED VALUES INDICATING DIFFERENCES BETWEEN SCHOOLS IN THE DISTRIBUTION OF SCORES, FOR TEACHER SELECTED PUPILS, ACROSS THE THREE SCORE CATEGORIES DEFINED IN THE TEST

Variable	English	Schools biased towards scores		French	Schools biased towards scores		Physics	Schools biased towards scores		Maths	Schools biased towards scores		df
		low	high		low	high		low	high		low	high	
FLU	12.8			8.4			11.2			7.1			12^{3}
FLEX	32.5			22.1			23.5			20.8			24
ORIG	12.1			10.2			14.1			10.2			12^{3}
A	36.6*	05,11	00	43.3**	05,11	04,07	41.3*	01,11	04	28.6			24
B	2.41			1.59			1.57			1.61			12^{2}
C	2.89			5.10			2.53			2.44			12^{2}
D	5.76			14.94			22.75*	03,07 12	02,09	18.40			12^{3}
E	26.07*	02,06	07,11	26.97**	06	02,07	22.45*	06	07	28.82*	06	07,11 12	12^{3}
F	14.98			21.28*	02	06	16.12			15.91			12^{3}
G	12.62			10.71			9.91			10.43			12^{3}
I	23.27	07,01		9.69			24.43*	03		14.04			12^{3}
VR	2.59			3.79			2.33			1.25			12^{2}
NA	1.26			2.58			1.70			1.84			12^{2}
VN	2.73			5.34			2.14			1.72			12^{2}
AR	0.85			2.30			2.04			2.08			12^{2}
CSA	3.06			4.17			3.61			4.40			12^{2}
MR	1.53			2.59			1.31			1.36			12^{2}
SR	1.96			2.06			1.54			0.43			12^{2}
SP	0.83			1.94			1.64			0.71			12^{2}
LD	3.69			3.26			2.92			1.91			
SEX[1]	1.71			16.23**	00,13 more boys 03 more girls		9.03			3.67			10
NEATNESS	28.60**	01,02 09	05	28.80**	07	05,04	17.00			22.10*		07,12	12
SOCIAL CLASS	20.97			18.29			19.51			13.49			12

1 Natural categories were used for these variables.

2 Small number problems for this comparison were solved by combining the categories 0 to 50th and 50th to 84th percentile.

3 Small number problems for this comparison were solved by combining the categories 50th to 84th and 84th to 100th percentile.

4 Significances are shown where they reached 5% level. * - 5% ** - 1%

Differences between groups of schools

Having explored individual school differences as reported above, we turned our attentions to groups of schools and sought to establish whether there were differences between, say, 11+ entry and 13+ entry schools with respect to the score distributions of the teacher-selected pupils.

We first give an example of the procedure used. When we tackled the question of difference between 11+ entry and 13+ entry schools in terms of the score distributions of teacher-selected pupils on, for example, the Flexibility test, we effectively went back to Table 4.10 and, for each subject, totalled up the numbers of pupils in each of the score ranges in the 11+ entry schools (00, 01, 03, 04, 06, 07, 08, 10, 11, 12) and in the 13+ entry schools (02, 05, 09). We then compared these two overall distributions by means of the chi-squared test.

Tables such as Table 4.11 could be presented for each grouping of schools studied (11+ v 13+; large v medium v small) but to do so would be to overload the reader unnecessarily, for very few significant differences were found. Instead we note in Table 4.12 the differences which were significant.

TABLE 4.12

SCORE DISTRIBUTIONS IN THE TEACHER SELECTED GROUPS
COMPARISONS BETWEEN GROUPS OF SCHOOLS

GROUPS COMPARED	Significantly different distributions in the teacher selected groups in:			
	ENGLISH	FRENCH	PHYSICS	MATHS
11+ entry v 13+ entry	CSA more high scoring pupils in 11+ schools X^2 = 12.95 df = 2,**	Attitude A X^2 = 7.94, df = 2,* more high scoring pupils in 11+ schools CSA, more high scoring pupils in 11+ schools X^2 = 9.30 df = 2,**	-	CSA, more high scoring pupils in 11+ schools X^2 = 7.29 df = 2,*
11+ schools only Large v Medium v Small	Attitude A① X^2 = 16.3 df = 4,** Attitude F② X^2 = 10.48 df = 4,* Attitude I X^2 = 11.90 df = 4,** Mid. sized schools have fewer low scoring pupils	Attitude A① X^2 = 14.0 df = 4,**	Attitude F② X^2 = 10.23 df = 4,*	Attitude A① X^2 = 17.1 df = 4,** Attitude F② X^2 = 14.51 df = 4,**
Control v Experimental schools	-	-	Attitude I X^2 = 9.42 df = 2,* Control schools have fewer low scoring pupils IQ X^2 = 10.55 df = 2,** Control schools have more high IQ pupils	Neatness X^2 = 6.42 df = 1,* Expt. schools have more high scoring pupils

① In English, French and Maths the large school had more low scoring pupils in its teacher selected groups. The mid sized schools had more high scoring pupils.

② Large school: more high scoring pupils; Mid sized schools: more mid scoring pupils; Small schools: more low scoring pupils.

Characteristics of teacher-selected pupils compared with pupils from the B Group who were not selected

As mentioned earlier, the B Group represents a fairly homogeneous set of pupils all of whom could be said, once due note is taken of the error in the regression equations, to have a 'good chance' of achieving an actual O-level score which would place them in the top 10% category.

In seeking characteristics which distinguish the pupils in the B Group who were selected, from those who were not, we hope to bring more strongly into focus the nature of variables which, consciously or unconsciously influenced teacher choice. That this is a reasonable hope can be seen from the following model of the selection process.

When teachers selected pupils for the top 10% group, some were included and some disregarded with some certainty. There was then a fairly large group of pupils, all of whom were quite good at the subject, but about each of whom the teacher was to some degree uncertain. The tests enabled us to identify the B Group as a group of pupils who would fall into this 'uncertain' group.

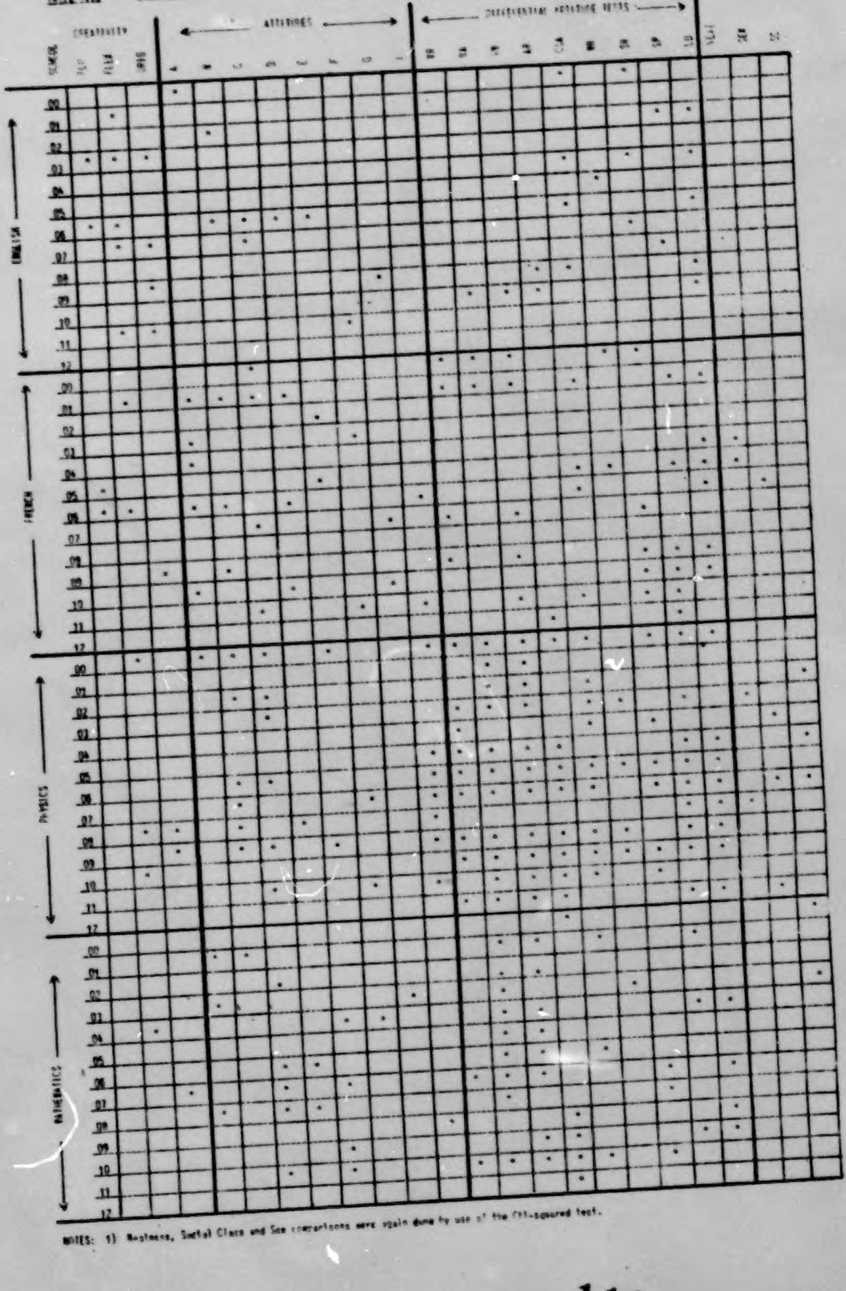
In looking for distinguishing features of the pupils who were finally selected from this group, we should be able to see the things (perhaps neatness or spelling or social class) which tip the balance of teachers' judgement in favour of including the pupil in the top 10% group.

The fact that some B Group pupils may have been misplaced by the schools setting procedure (and might not therefore be in the group from which the teacher is making his or her selections) suggests that this analysis of the B Group may reflect on the nature of the overall selection process in the school, rather than just on the process underlying the final selection of the top 10%.

Finally, as preamble to the discussion of the results of this analysis, it is interesting to reflect that, if the teacher-selected group in a given subject is indeed characterised by high or low scores on some variable, then these characteristic variables should show up in comparisons with any other group in that school year. The B Group merely represents a group that will show up the character of selection because it is a group where anyone would have been a 'good' choice. We will have some clues therefore to the kind of 'fine tune' of the selection process that may be needed.

The method of analysis via correlation and partial correlation was the same as the method used for the study of the characteristics of teacher-selected pupils in relation to the whole of the rest of the year group. This analysis was supported by other statistics and details of these are again given in Appendix 4. The most useful results are shown below.

As mentioned in the introduction to this chapter, this part of the analysis is likely to bring to light any rather inappropriate biases which affected teachers' selection procedures. Because the range of scores on test scales that were appropriate to the measurement of potential in the subject is likely to be small within the B Group, we would only expect to find differences between selected and non-selected pupils on these scales if the teachers were particularly sensitive to the assessment of the abilities that they measure. We expect, therefore, to find that the results of this part of the analysis will be dominated by any rather less relevant characteristics that encourage teachers to distinguish the pupils whom they selected, from others of fairly similar ability. Tables 4.13 and 4.14 summarise the results of this work by showing the variables which were found to be significantly correlated with teacher selection (at or above the 5% level of significance).


CREATIVITY
SCHOOL
ATTITUDES
A
B
C
D
E
F
G
H
DIFFERENTIAL APTITUDE TESTS
ENGLISH
FRENCH
PHYSICS
MATHEMATICS
00
01
02
03
04
05
06
07
08
09
10
11
12

NOTES: 1) Neatness, Social Class and Sex comparisons were again done by use of the Chi-squared test.

TABLE 4.14

NO. OF SCHOOLS SHOWING SIGNIFICANT CORRELATION BETWEEN SELECTION AS 'MORE ABLE' AND EACH TEST SCALE WITHIN THE 'B GROUP' (ROUND 1)

	ENGLISH	FRENCH	PHYSICS	MATHS
CREATIVITY				
FLU	2	2	0	0
FLEX	3	2	3	1
ORIG	4	1	2	1
ATTITUDES				
A	1	5	1	3
B	2	3	6	1
C	2	4	6	5
D	1	3	1	2
E	1	2	2	4
F	1	2	2	1
G	1	2	0	1
I	0	2	7	1
DAT				
VR	0	4	9	2
NA	1	2	11	9
VN	1	5	11	8
AR	2	1	8	5
CSA	4	4	9	3
MR	1	2	4	1
SR	3	5	6	3
SP	2	6	9	3
LU	5	6	8	4
SEX	0	2	2	0
NEATNESS(NT)	0	1	3	0
SOCIAL CALSS(SC)	0	0	3	2

A scrutiny of the tables indicates that, in general, fewer significant differences were apparent, though in physics there was a tendency for the selected group to be higher in ability on a large number of DAT scales. 'Favourable' attitudes featured as characteristics of the teacher-selected groups in most schools and high creativity was also a characteristic of some groups. Few sex, social class and neatness differences showed up. These variables seemed to have most impact in physics, where the results of even this fairly insensitive study of the characteristics leads one to suspect that the most appropriate characteristics were not always present. A summary of these results is useful and one sees that variation from school to school was a feature of the characteristics.

An attempt is made later to assess the relevance of the characteristics but it is interesting to observe how many schools characterised their selection on the 'most appropriate' scales, even though it was difficult for teachers to be sufficiently sensitive to these abilities for the characteristics to show up in this particular analysis. About half of the schools, in both English and French showed a link with LU (the best single predictor for these subjects). Nine out of the thirteen schools show NA as a strong characteristic of the mathematics selection. However, only four schools showed MR as a characteristic in the physics selection. Instead, VN was a strong characteristic which emerged in as many as eleven schools in physics. VN did not seem to be so dominant in other subjects. It emerged in only one school in English and in two schools in French.

Partial Correlation, again, allows us to identify the strongest characteristics of the selected groups. Table 4.15 shows these results. The main implication is that individual teachers displayed individual biases. However there was a consistent tendancy for selected pupils in physics to have a high VN score (so that the apparently strong influence of this variable to which we have already drawn attention, cannot be explained away in terms of correlations between VN and other variables). In the almost total absence of high MR as a characteristic of selected pupils in this table (and even more significantly in Table 4.4) it is tempting to infer that physics teachers, either at the stage of drawing up science sets or at the stage of final selection for the 10% group (or as a consequence of working with streamed rather than setted classes) did seem to place too much emphasis on general ability and to be insufficiently aware of specific scientific abilties. (Some might argue that such abilities have not developed by the third year. The fact that the regression equations picked out MR rather than the general ability VN as the best predictor for physics O-level makes this a difficult point to sustain.) It is also interesting to note that high Clerical Speed and Accuracy (CSA) or high Spelling (SP) were characteristic of selection in three schools in mathematics, in four schools in English (where admittedly SP would not be an unreasonable characteristic), in four schools in French and in no less than six schools in physics.

We doubt whether the characteristics displayed show the full picture. The conscious, or subconscious process of selection may take account of qualities of individuals that our tests do not measure, or that are only correlated with the tests that appear in Table 4.15. On the other hand, it may be that the varied picture reflects the true difficulty of assessing who is more able when one must select, say, the top half of the top set. Further attention is given to the question of whether the appropriate characteristics of ability were picked up by teachers later in the third year in Chapter Six, where we examine the Round 2 results. At that stage we will also see if checklists sharpen up the ability to pick out these appropriate characteristics.

TABLE 4.15

MAIN VARIABLES CHARACTERISING THOSE CHOSEN IN 'B GROUP' FROM PARTIAL CORRELATION TECHNIQUE

SCHOOL	ENGLISH	FRENCH	PHYSICS	MATHS
00	CSA	(VR and VN)	CSA or C	AR
01	-	(CSA and A)	(NA and VN)	(CSA and A)
02	B^1 or SP or LU	-	(B and NA)	(C and NA)
03	(FLU and FLEX and ORIG)	A	VN or SP or LU	A or NA
04	LU	A	CSA or VR	(B and NA)
05	-	E^1 or LU	SP or VN	NA or VN
06	LU or B, or C or D	LU or (A and B)	B or VN	C or VN
07	FLEX or SR	VN or SC	SR or (F and I)	SR
08	SP	-	FLEX or SP	CSA or (A and D)
09	LU or CSA,^{1}or ORIG of G^1	CSA or SP	VN or C	AR or LU
10	(VN and AR and NA)	A	VN or MR or LU	VN or AR
11	(ORIG and FLEX)	C or CSA	SR or CSA or C	(NA and VN) or AR or CSA
12	-	SP	LU	AR

Accuracy of B Group selection

It is possible, to some extent, to judge the success of a particular school in selecting the most appropriate pupils in a given subject by judging whether, in the 'B Group', the 'better' candidates were selected. If the ones who, in general, had higher predicted scores were being selected, evidence of an appropriate selection strategy would be seen. We tackled this in two ways. First we found the point biserial correlations of teacher selection with predicted score, and secondly we compared the mean of the teacher selected group with that of the rest of the 'B' Group by using the *t* test. Table 4.16 summarises the results for those schools and subjects which yielded significant correlation or significant t value. The strategies are again difficult to analyse and give an indication that more work is needed to establish the 'real' factors which were at play in the identification process. It is interesting, however, that where significance is at the 0.1% (***) level[9], the characteristics do seem appropriate to a good identification strategy.

TABLE 4.16

SCHOOLS AND SUBJECTS SHOWING 'GOOD SELECTION STRATEGY' IN B GROUP WHEN THE PREDICTED SCORE ON THOSE SELECTED IS COMPARED WITH THOSE NOT SELECTED

SCHOOL	SUBJECT	MAIN VARIABLES USED	CORRELATION BETWEEN SELECTION AND PREDICTED SCORE			t TEST[1] FOR SIGNIFICANT DIFFERENCE IN MEAN			
			r	N	SIG	t	N_1	N_2	SIG
01	MATHS	(CSA AND A)	.3541	35	*	2.18	12	23	*
02	PHYSICS	(B AND NA)	.5447	36	***	3.79	12	24	***
02	MATHS	(C AND NA)	.4637	39	***	3.18	15	24	**
03	ENGLISH	(FLU AND FLEX AND ORIG)	.5026	26	**	2.85	7	19	**
03	PHYSICS	(VN OR SP OR LU)	.3471	33	*	2.06	9	24	*
04	ENGLISH	LU	.2940	34	*	-	-	-	NSD
05	FRENCH	(E OR LU)	.4038	19	*	-	-	-	NSD
05	PHYSICS	(VN OR SP)	.4216	22	*	2.08	5	17	*
05	MATHS	(NA OR VN)	.5115	19	*	2.45	7	12	*
06	ENGLISH	LU OR B, C, D OR G	.6095	39	***	4.68	11	28	***
07	PHYSICS	SR OR (F AND I)	.4386	36	**	2.85	4	32	**
07	MATHS	SR	.4184	36	**	2.69	14	22	*
08	ENGLISH	SP	.3740	28	*	2.06	9	19	*
08	FRENCH	-	.3985	18	*	-	-	-	NSD
09	ENGLISH	LU OR CSA OR ORIG OR G	.3807	38	**	2.45	14	24	*
09	PHYSICS	VN OR C	.3502	53	**	2.67	14	39	*
09	MATHS	AR OR LU	.4260	46	**	3.13	17	29	**
10	PHYSICS	VN OR MR OR LU	.6078	22	***				
10	MATHS	VN OR AR	.4170	19	*				
11	MATHS	(NA AND VN) OR AR OR CSA	.5988	21	**				
12	PHYSICS	LU	.6816	17	***				
12	MATHS	AR	.4045	19	*				
13[2]	PHYSICS					2.59	17	50	*
13	MATHS					4.7	24	40	***

all cases the F value showed no significant difference in variance for results shown in this table.

ecall that schools 10, 11 and 12 are halls of school 13. The T tests on school 13 Physics and Maths confirm the ttern shown in the three halls. The confined result allowed larger numbers for the t test.

If one tries to identify the type of school that produced 'good' strategies in a reasonable number of subjects there is evidence that schools 02, 03, 05, 08, 09 had 'good' strategies on at least two subjects. This group of schools includes all of the 13+ entry schools and the small 11+ entry schools. The largest school also had 'good' strategies on two subjects, but of course it can be thought of as three halls of one school - three 'small' units. Of the five 'medium' sized schools (1000-1500 pupils), only three (01, 04, 07) appeared in Table 4.16 and only one of these (07) appeared twice. One might tentatively suggest that 13+ entry schools, or that small schools (or schools which operate as small subsections) produce more effective identification strategies. However, it is not possible to decide clearly between these alternatives on the basis of this evidence alone.

B Group pupils: Attitudes to the subjects

Chi-squared tests (with Yates Correction where appropriate) or Fisher tests were carried out for the B Group, as they were in the previous work when teacher-selected groups were compared with the rest of the year group, to examine the relationship between teacher selection and i) pupils likes/dislikes, ii) their views of the usefulness of a subject and iii) their perception of their own ability in the subject.

Fewer significant differences were found than in the whole year group comparison. Table 4.17 lists those cases where a significant difference was found.

TABLE 4.17

THOSE SCHOOLS AND SUBJECTS THAT SHOWED A SIGNIFICANT DIFFERENCE ON ATTITUDE TO THE SUBJECT WHEN THOSE PUPILS WHO WERE SELECTED WERE COMPARED WITH THOSE NOT WITHIN 'B GROUP' (ROUND 1)

SCHOOL	'LIKE'	'DISLIKE'	'USEFUL'	'NOT USEFUL'	'GOOD AT'
ENGLISH					
09	*				
FRENCH					
04			*		
06	**	*			
09	*	*			*
10					*
11	*				
12					*
PHYSICS					
02			*		
03					**
MATHS					
02					***
03					*
12	*[1]				

---

[1] This difference unlike all other indicated a stronger liking in the non-selected group.

If anything French identification seemed more sensitive than any other subject, but school by school, these factors did not strongly characterise teacher selection.

Table 4.18 is the summary table which gives the totals when the figures for individual schools were pooled. Table 4.19 shows where, in Table 4.18, there were significant differences between selected and non-selected B Group pupils. We can use this table to see the nature of any consistent trend across all schools.

TABLE 4.18

COMPARISON OF THOSE IDENTIFIED WITH THOSE NOT WITHIN 'B GROUP' ON ATTITUDES TO THE SUBJECT (ROUND 1)

	SAID 'LIKE'	NOT	SAID 'DISLIKE'	NOT	SAID 'USEFUL'	NOT	SAID 'NOT USEFUL'	NOT	SAID 'GOOD AT'	NOT
ENGLISH	92	30	4	118	104	18	1	121	77	45
	188	74	32	230	224	38	3	259	122	140
FRENCH	56	25	13	68	45	36	6	75	40	41
	66	121	79	108	64	123	29	158	48	139
PHYSICS	62	50	22	90	61	51	14	97	43	68
	172	165	82	255	172	165	28	309	99	238
MATHS	89	44	17	116	123	10	0	133	78	55
	153	98	48	203	225	26	5	246	121	130

TABLE 4.19

TABLE SHOWING WHICH RESULTS ARE SIGNIFICANT FROM TABLE[1]

	'LIKE'	'DISLIKE'	'USEFUL'	'NOT USEFUL'	'GOOD AT'
ENGLISH	NSD	**	NSD	NSD	**
FRENCH	***	***	**	NSD	***
PHYSICS	NSD	NSD	NSD	NSD	NSD
MATHS	NSD	NSD	NSD	NSD	NSD

---

[1] Chi-squared tests were used or Fisher Exact Probability Test when numbers were too-small for Chi-squared.

It is useful to look at these results subject by subject. In English there is an indication that thinking oneself good at the subject, and not positively disliking the subject, were characteristics which distinguished between the selected and the non-selected B Group pupils. However, there was a large proportion of pupils (25% of all pupils selected by their teachers) who were identified and yet who did not say that they liked the subject. More than 30% who were chosen did not think themselves good at the subject. The lack of significant difference on usefulness of subject reflects the overall opinion of both those selected and those not selected that English is a useful subject.

In French, there is a clear indication that in this overall comparison, attitudes towards the subject played some part in the identification proceess. Nevertheless a fairly high proportion of those identified (approximately 30%) were not prepared to say they liked the subject, approximately 45% were not prepared to say the subject was useful and only 50% thought that they were good at the subject. The significances reflect the fact that the situation was much more extreme for those not chosen.

In physics no evidence was found that the attitudes affected selection, but a large percentage (approximately 40%) did not say that they thought themselves good at the subject. This may be because they had, in general, not experienced physics as a subject for very long.

In mathematics a large number (33%) of those identified did not say they liked the subject and 12% actually disliked the subject. Over 40% of those selected did not say that they thought themselves good at it and this reflects the uncertainty of ability across all the subjects. Very few in the whole B group thought that the subject was not useful, however, and there was no significant difference between those selected and those not.

Trends across the sample as a whole

An important study is that of the biases that can be detected in the trends over the sample as a whole: The first part of this work followed a previous pattern, and compared the teacher-selected pupils with the whole of the rest of the year group. By grouping all schools together we hoped to reveal any consistent trends for teachers to select, say, boys or high social class pupils. In view of the large number of significances found in the previous work on individual schools, we can assume that there would be no new insights except in the case of the three categorical variables.

Table 4.1 gives some hints as to what might be expected. We therefore set out the results for neatness, sex and social class separately below, referring to this table as appropriate.

High neatness score seemed, from Table 4.1, to distinguish the teacher-selected pupils from the rest of the year group in several schools in the subject areas of English and French. We might therefore expect the sample as a whole to reflect this characteristic. However in mathematics and physics there were few significant differences for individual schools. It is, of course, impossible to tell from this whether or not there was an underlying trend in all schools which just fails to reach significance in each one. We therefore found the total number of teacher-selected pupils, across all schools, who scored 1, 2, 3, or 4 or 5 on the neatness measure. (Scores of 4 and 5 were put together as very few scores of 5 were achieved.) We then compared this distribution with the similar distribution for the non-selected pupils.

Table 4.20 indicates this procedure for the subject of English to serve as an example. Table 4.21 summarises similar work in the other subject areas.

TABLE 4.20

COMPARISON OF THE NEATNESS OF TEACHER SELECTED PUPILS
WITH THE REST OF THE YEAR GROUP.
ALL SCHOOLS TOGETHER.
ENGLISH

	Number of pupils with neatness scores of:								
	Observed Table				Expected Table				Chi-squared result
	1	2	3	4 or 5	1	2	3	4 or 5	
TEACHER SELECTED PUPILS	11	72	94	36	27.8	93.9	77.7	13.6	$\chi^2 = 61.80$ df = 3 $p \doteq 0.001$
NON-SELECTED PUPILS	256	830	652	95	239.2	808.1	668.3	117.4	

We can see from the observed and expected tables that the teacher-selected group contained more neat pupils than we would have expected if teacher-selected and non-selected pupils were equivalent on this measure. We see from the chi-squared value that this difference is very unlikely (probability $\doteq$ 0.001) to occur by chance.

We conclude that an overall characteristic of the teacher-selected pupils in English across the whole sample was that they were neater than the non-selected pupils.

TABLE 4.21

SUMMARY OF DIFFERENCES IN NEATNESS FOR ALL SUBJECTS

	ENGLISH	FRENCH	PHYSICS	MATHS
X^2, df = 3	61.80	64.91	17.87	21.16
Significance	$p=0.001$	$p=0.001$	$p=0.001$	$p=0.001$

All of these differences were consistent with the result in English, namely: that the teacher-selected groups were neater than the rest. It is interesting to see not only confirmation of our suspicions that there would be such a difference in English and French, but also a new finding that a significant overall trend could be detected in physics and mathematics.

We conclude that although few individual schools demonstrated a significant difference on this scale, there was nevertheless an underlying trend which, accumulated across the sample as a whole, was revealed as significant.

If we felt that this 'high neatness' characteristic of teacher-selected pupils was an error (and we must of course be careful in making this judgement because, as explained before, this characteristic might arise simply because neatness is related to other more appropriate scales) then in physics and mathematics we might expect it to be particularly difficult for teachers to correct this error as it is only across the sample as a whole that it becomes very apparent. Individual teachers might therefore get relatively little feedback about the existence of this characteristic in their own selection.

Similar work on sex is summarised in Table 4.22.

TABLE 4.22

SUMMARY OF SEX DIFFERENCES BETWEEN TEACHER SELECTED GROUPS AND NON SELECTED GROUPS FOR THE SAMPLE AS A WHOLE

	ENGLISH	FRENCH	PHYSICS	MATHS
X^2 (df = 1)	20.99	36.74	1.05	0.87
Significance	$p \leq 0.001$	$p \leq 0.001$	nsd	nsd

In English and French the teacher-selected group contained a higher proportion of girls than the rest of the sample. It is interesting to note that the indications from Table 4.1 that there would be a sex bias in French were confirmed and that in addition an overall trend was revealed in English, but not in physics or mathematics.

Similar work on social class is summarised in Table 4.23.

TABLE 4.23

SUMMARY OF SOCIAL CLASS DIFFERENCES BETWEEN TEACHER SELECTED GROUPS AND NON SELECTED GROUPS FOR THE SAMPLE AS A WHOLE

	ENGLISH	FRENCH	PHYSICS	MATHS
X^2 (df = 3)	88.17	89.49	92.35	68.17
Significance	$p \leq 0.001$	$p \leq 0.001$	$p \leq 0.001$	$p \leq 0.001$

In all subjects the teacher-selected groups contained higher proportions of pupils from social classes 1 and 2 (the professional end of the scale). These results are not surprising in the light of the results of Table 4.23.

The analyses described above were also carried out comparing, for all schools taken together, the teacher-selected pupils with the other pupils from the A and B Groups only who were not selected. This work extends previous insights gained for this group and indicates overall characteristics which were associated with selection not, as above, from the whole of the schools' year group, but from pupils within those year groups who have high, (and roughly comparable) ability in each of the subjects.

Some fascinating results were forthcoming. On the neatness scale, the teacher-selected group was significantly more neat than the group of non-selected A and B Group pupils in French and in physics only. Therefore, although English and mathematics teachers seemed to have selected pupils who were the neater ones in the school as a whole (see Table 4.21), they did not seem to select the neatest of the able pupils in their subject. One could interpret this as showing that they were not unduly influenced by neatness, and that the association between selection and neatness for the year groups as a whole was simply due to a correlation between neatness and the scales on which teachers in English and mathematics were actually making their selection.

However, in French and physics there was a tendency for teachers to select the neatest pupils from the able group. Perhaps, then, neatness was given too much weight by teachers of these subjects. This is an intriguing conclusion in the light of the generally lower accuracy of teachers in French and physics which we have reported elsewhere.

The social class analysis indicated significant differences between teacher-selected and non-selected A and B Group pupils in all subjects but French. This might, in a similar way to that described above, imply that teachers were being unduly influenced by social class.

The analysis in terms of sex raised an interesting finding. In all four subjects the teacher groups were found to contain higher proportions of girls than were present in the rest of the A and B Groups. This can be interpreted as showing that, in physics and mathematics (where no overall bias in favour of girls was apparent from comparisons with the rest of the whole year group) there were actually more able boys than there were able girls, and that the more or less equal proportions of boys and girls selected by teachers still therefore revealed a tendency to pick girls at the expense of equally able boys. The finding that there were more able boys than there were able girls in physics and mathematics is consistent with the findings of Astin (1975) who studied the relationship between sex and ability in these subjects.

Table 4.24 summarises the chi-squared analysis on which these comments are based.

TABLE 4.24

SUMMARY OF DIFFERENCES BETWEEN TEACHER SELECTED PUPILS AND OTHERS IN THE A AND B GROUUPS COMBINED

	ENGLISH	FRENCH	PHYSICS	MATHS
NEATNESS[1]				
X^2 (df = 3)	2.91	13.31	11.02	3.74
Significance	nsd	$p \leq 0.001$	$p \leq 0.005$	nsd
SOCIAL CLASS[2]				
X^2 (df = 3)	9.26	6.20	28.12	16.46
Significance	$p \leq 0.05$	nsd	$p \leq 0.001$	$p \leq 0.001$
SEX[3]				
X^2 (df = 1)	5.99	13.22	15.08	4.98
Significance	$p \leq 0.05$	$p \leq 0.01$	$p \leq 0.01$	$p \leq 0.05$

1. The teacher selected groups contained proportionately more pupils with higher neatness scores of 3, 4 or 5.

2. The teacher selected groups contained proportionately more pupils in Social Class 1 or 2.

3. The teacher selected groups contained proportionately more girls.

It ought, perhaps to be emphasised that these results only emerged from the analysis when the results for the individual schools were pooled. They rarely revealed themselves within an individual school (see Tables 4.4, 4.14 & 4.15). These are not inconsistent findings. They merely emphasise that there are slight but consistent trends in the way in which the teachers in the whole of our sample responded to the three variables that we have been discussing.

Characteristics of the A Group pupils who were selected by teachers

We recall that conclusions about the A Group are conclusions about roughly the top 2% on test ability. As we showed in Chapter Three, not all of the A Group pupils were identified by teachers. We attempted to shed some light on the causes of the disagreement between the teachers and the tests over the ability of these pupils by trying to determine what set the pupils who *were* selected by teachers apart from those who were also in the A Group but who were *not* selected by teachers.

As an example of the first part of this analysis we set out below the comparison of the selected and non-selected pupils in terms of sex. Table 4.25 shows the numbers of boys and girls from the A Group who were and who were not selected by teachers in the subject area of English. The figures are for all schools taken together.

TABLE 4.25

TEACHER SELECTED AND NON-SELECTED A GROUP PUPILS CATEGORISED BY SEX

	GIRLS	BOYS
No of A Group pupils:		
selected by teachers	39	25
not selected by teachers	9	6

When this table is subjected to chi-squared analysis using Yates correction we find a chi-squared value of 0.05 (df = 1) which indicates that there was no significant difference between the teacher selected and the non-selected pupils of the English A Group, in terms of distribution between the sexes.

Similar analysis for the other subjects produced the similar non significant values of chi-squared listed below.

TABLE 4.26

SEX DIFFERENCES BETWEEN TEACHER SELECTED AND NON SELECTED MEMBERS OF THE A GROUP

	ENGLISH	FRENCH	PHYSICS	MATHS
X^2 (df = 1)	0.05	0.00	0.89	2.88
Significance	nsd	nsd	nsd	nsd

Similar analyses were then undertaken to search for differences in terms of pupils' neatness and social class. Again no significant differences were discovered. In order to discover differences revealed by the other background tests the mean scores of those chosen were compared with the mean scores of those not by the t-test. There was not a single significant difference in any subject.

Finally comparisons were made between the A Group pupils who were selected by teachers and those who were not, in terms of their attitudes, not to school generally, but to the particular subject under discussion. For example we found that 85% of the A Group pupils who liked English were identified by teachers whereas only 50% of those in the A Group who did not like English were identified.

Table 4.27 shows similar figures for three aspects of A Group pupils' views of the four subjects.

TABLE 4.27

RELATIONSHIP BETWEEN PUPILS' VIEWS OF A SUBJECT AND THEIR IDENTIFICATION AS MORE ABLE BY THEIR TEACHERS (A GROUP)

	PUPILS' VIEWS:					
	'Like'		'Subject will be useful'		'Subject is one I am good at'	
	No	Yes	No	Yes	No	Yes
TSEL*	5	58	9	54	21	42
ENGLISH						
NOT**	5	10	2	13	8	7
Chi-squared (sig)	6.99 (1%)		0.00 (nsd)		2.07 (nsd)	
TSEL	18	36	18	36	22	32
FRENCH						
NOT	20	14	22	12	24	10
Chi-squared (sig)	5.52 (5%)		8.28 (1%)		7.45 (1%)	
TSEL	8	29	10	37	17	30
PHYSICS						
NOT	12	19	12	19	19	12
Chi-squared (sig)	2.37 (nsd)		2.80 (nsd)		4.74 (5%)	
TSEL	8	39	5	32	7	30
MATHS						
NOT	5	10	1	14	4	11
Chi-squared (sig)	16.60 (0.1%)		0.49 (nsd)		0.38 (nsd)	

* TSEL = Selected by teacher
** NOT = Not selected by teacher

It would seem that, in English, French and mathematics teacher selection of pupils from the A Group was linked with whether or not the pupils liked the subject (though of course it is dangerous to infer causality: maybe teachers selected pupils who liked the subject; maybe pupils who were well regarded by their teachers grew to like the subject). Similarly, in French and physics, teacher selection of pupils was linked to the pupils' own views of their ability in the subject. In French teacher-selected pupils also tended to be those who thought that the subject would be useful.

In a separate set of questions pupils were also encourged to state whether they actively dislíked a subject or felt it definitely to be not useful. In French and physics pupils feeling that the subjects would not be useful tended not to be selected by teachers. In French pupils actively disliking the subject tended not to be selected.

These aspects of pupils' feelings about a subject were therefore found to be associated with whether or not they were selected as 'top 10% pupils' by their teachers but it is perhaps important to repeat that it is dangerous to infer that these pupil characteristics were the reason for the teacher selection (or non selection).

As well as the overall differences outlined above individual school differences were explored. In all cases schools showed trends which were in agreement with the overall results. In a few cases these were strong enough to be statistically significant within the one school.

Characteristics of C Group pupils who were identified by teachers

A similar analysis to that undertaken for the A Group was conducted in an attempt to discover variables on which the teacher-selected C Group pupils could be distinguised from the rest of the C Group. In this, we were seeking clues as to why pupils whose test scores indicated low potential were nominated as able. The analysis was exactly the same as for the A Group, and we looked for trends across all schools. The only subject to show a difference on sex was French where there was a large bias towards choice of girls. Table 4.28 shows the figures for this comparison.

TABLE 4.28

TEACHER SELECTED AND NOT SELECTED C GROUP PUPILS CATEGORISED BY SEX

	GIRLS	BOYS
Selected by teachers	50	14
Not selected by teachers	601	758

Chi-squared = 26.95 $p \leq 0.1\%$

The other categorical variables (social class and neatness) gave no significant differences in any subject. The t-test comparisons for the other background scales revealed an assortment of significances as shown in

Table 4.29.

TABLE 4.29

SHOWING THOSE VARIABLES IN EACH SUBJECT THAT
CHARACTERISED SELECTION BY TEACHERS IN THE
C GROUP AS INDICATED BY A t-TEST SIGNIFICANCE OF 5% OR LESS

ENGLISH	NA								
FRENCH	ORIG	A	B	VR	NA	VN	SR	SP	LU
PHYSICS	A	B	C	VR	NA	VN	SR	SP	LU
MATHS	B	C	NA	VN					

The intercorrelations between these variables are difficult to untangle, but there does seem to be a bias towards the variables that make up the general ability scale VN (VR and NA). This was particularly so in French and physics. In English there was no dominance of the most appropriate variable (LU) and, likewise, in physics the most appropriate variable (MR) was not present in the table. Table 4.30 gives the results of chi-squared analyses which were made in connection with attitude variables.

TABLE 4.30

CHI-SQUARED VALUES FOR THE COMPARISON
OF C GROUP PUPILS SELECTED BY TEACHERS AND THOSE
NOT SELECTED IN TERMS OF PUPILS' FEELINGS ABOUT THE SUBJECT

	PUPILS' VIEWS:				
	'Like'	'Dislike'	'Will be Useful'	'Will not be Useful'	'Subject is one I am good at'
ENGLISH	0.25 (nsd)	2.60 (nsd)	2.41 (nsd)	0.92 (nsd)	35.99 (0.1%)
FRENCH	59.43 (0.1%)	38.85 (0.1%)	44.74 (0.1%)	20.67 (0.1%)	74.15 (0.1%)
PHYSICS	20.60 (0.1%)	8.81 (1%)	26.81 (0.1%)	2.29 (nsd)	28.91 (0.1%)
MATHS	21.71 (0.1%)	8.70 (1%)	4.96 (5%)	0.89 (nsd)	36.13 (0.1%)

In English, teacher selection was linked strongly with the pupils' view of their ability in the subject. This was also an important factor in all other subjects.

In all subjects the proportion of pupils who thought they were good at the subject was higher in the teacher-selected group than in the non-selected group.

In all subjects except English the scales 'I like the subject' and 'The subject will be useful' were related to teacher selection in a similar way.

Pupils who actively disliked these subjects were less likely to be selected by teachers.

Concluding comments

It is extremely difficult to sum up the detailed findings of this chapter in a succinct and readable way. In very general terms there does seem to be an indication that physics teachers took too much account of general ability, and we do have some evidence that sex and social class, though they were to some extent related to selection in all subjects, had by no means as strong an influence on teachers' judgements as is sometimes supposed. Perhaps the most important generalisation which can be made is that it is dangerous to make generalisations about the nature of the teacher selection process. Different teachers seemed to be influenced by different characteristics. This finding emphasises the need for a study which is designed to be much more sensitive to these individual differences. What we discovered, as a result of our attempt to carry out such a study, are set out in later chapters of this report.

Notes for Chapter Four

1) Only trends across all schools could be measured because of the small numbers in the A Groups.

2) The technique used was as follows. Firstly, only those variables that were still significantly correlated with selection when scores were adjusted for all other variables are included in the list. If variable X is significantly correlated when controlling for Y, and Y is significantly correlated when controlling for X, then X *and* Y are characteristics. If X is not significantly correlated when controlling for Y, and Y is not significantly correlated when controlling for X, then X *or* Y are characteristics.

3) The questions asked were:

'Which of the following subjects do you like?'
'Which of the following subjects do you dislike?'
'Which of the following subjects will be most useful to *you* when you leave school?'
'Which of the following subjects will be least useful to *you* when you leave school?'
'Which of the subjects are you particularly good at?'

In all cases the four subjects, English, French, physics and mathematics were possible choices.

4) In cases where numbers were too small for a valid chi-squared test, Fisher's Exact Probability test was used instead.

5) School 13 was counted as 3 separate units for this analysis.

6) Any significance at or above the 5% level were counted. Many of these significances were above the 0.1% level.

7) The variables sex, neatness and social class which were already categorical variables, were not, of course, handled in this way. The distribution of pupil 'scores' across the natural categories of these scales was analysed by the Chi-squared test.

8) It is interesting to speculate that, in a school with a fairly 'difficult' intake the well adjusted pupil catches the teacher's eye

whereas in a more relaxed atmosphere teachers can more easily tolerate pupils who are not as well adjusted and are able to perceive them as able despite their other problems. Such speculation has, of course, to be treated with great care.

9) School 02 physics (B and NA), School 02 mathematics (C and NA), School 06 English (LU or B,C,D or G), School 10 physics (VN or MR or LU) and School 12 physics (LU).

CHAPTER FIVE

Preliminary Work on Checklist-Aided Identification

5.1 Introduction

In Chapter One we discussed the idea that teachers might attempt to improve that part of their record keeping that is relevant to their identification of more able pupils, by compiling a checklist of characteristics which are related to high ability in their own subject. Such a list would be used to record the levels reached by pupils on each of the characteristics. It has often been suggested in the 'gifted literature' that the use of such checklists would lead to an improvement in teacher-based identification of able pupils.

Certainly the teachers who were involved in our project placed great emphasis on the use of their own records when they constructed their top 10% lists for us. This can be seen from Table 5.1 which summarises the methods used by them in their identification of the top 10% of third form pupils in their subject. Teachers from eleven schools contributed data to this table.

TABLE 5.1

INFORMATION USED BY TEACHERS IN THE IDENTIFICATION OF THE TOP 10% IN A SUBJECT

Information	Number of departments using this kind of information			
	ENGLISH	FRENCH	PHYSICS	MATHS
Teachers' own records	10	11	10	11
School exam results	3	8	4	9
School reports	1	1	0	1
Objective tests	3	1	7	2
'Other'				
Pupils' ability to reason	–	–	1	–
Opinions of teachers of other subjects	–	2	–	–
Subjective opinion of class work	2	–	4	1
Middle school reports	–	–	1	–

From the responses of the teachers recorded in this table it would appear that teachers' own records were, in all four subjects, the most common sources of data. Very little use was made of school reports and interestingly only two departments (both French) sought advice from teachers of other subjects. Physicists seemed to be more likely than other teachers to take note of class work, but it may be that in French and English subjective assessment of oral work in class forms part of the usual teacher record, and reference by the teacher to the pupils' oral ability may well have been implied by ticking the 'own record' box. In the light of these responses it does certainly seem to be plausible that a systematic means of adding further data to the teachers' own records could be effective in improving identification procedures.

It is also reasonable to suggest that there may be room for improvement in teachers' records. Limited time is available for a secondary school teacher to get to know the capabilities of an individual child in depth. Large class size, shortness of lessons, the small number of lessons that a teacher spends with a given class each week and the consequently large number of classes with which a teacher deals, syllabus constraints and other pressure on the teacher's attention during lessons (such as the routine administrative activities - such as handing out and checking in books, and sorting out difficulties with apparatus) all mediate against the maintenance of adequate records. Checklists may therefore serve a very useful purpose... though it must be recognised that the very reasons which suggest that the teacher might find it helpful to use a checklist to improve his or her records, also lead one to suspect that the teacher may not have the time to fill in a checklist very carefully.

It should also be remembered that the construction of checklists is not without some difficulties. In Chapter One we have drawn attention to the problems of designing valid items for inclusion in a subject specific checklist. We would add, here, a further concern relating to the use of checklists: it is quite possible that, even if the items on the list are of proven validity, the pupils may not give sufficient clues to allow accurate judgements to be made of the levels which they reach on each item. This could come about if the kind of work done in class limits the kinds of abilities which a pupil could demonstrate. (There is an interesting possibility raised by this consideration: it may be that would be easier to identify able pupils more accurately if some provision, aimed specifically at such pupils, was included in normal classwork. This possible connection between provision and identification will be taken up in later chapters.)

We have already mentioned in Chapter One that many educationalists have recommended the use of checklists as a means of improving teacher-based identification. The points which we have set out above indicate that checklists may well be useful, but that their effectiveness cannot be taken as a foregone conclusion. It is therefore vital to study the effectiveness of such instuments so that one can confirm, refine or reject a method that could become widely used.

5.2 Research method for the preliminary study of checklists

In this section we will discuss the structure of the first stage of the work with checklists. In Section 5.3 we will discuss the evolution of the subject specific checklists that we used.

We have described the general strategy of the work on checklists in Chapter Two. The schools were divided into control and experimental groups. Both groups were asked to make an identification in the Winter Term without the use of checklists. It is the outcome of this round of identification that we have been discussing in Chapters Three and Four. The control schools were also asked to make a second identification of their top 10% towards the end of the first year, again without checklist support. The experimental schools were asked to make a second identification at the same time, but in these schools teachers were asked to make use of checklists. These teachers were given the checklists and asked to fill them in over a relatively short period of time (usually two to three weeks). This simulated the exercise that would be carried out in schools where teachers were making use of published checklists in the manner that is usually recommended.

By comparing the accuracy and the characteristics of identification in the experimental and control schools during the first round of identification we were able to establish the equivalence of the two groups of schools (i.e. to ensure that the random allocation of schools to the two groups had been successful in producing groups which behaved in similar ways). By repeating the comparison for the second round of identification we were able to see whether checklists had enabled the experimental schools to improve in relation to the control schools. Of course both groups might be expected to show some improvement because the second round of identification was carried out later in the year, when teachers had had longer to get to know their pupils. However the comparison of experimental and control schools enabled us to assess the effectiveness of the checklist approach, controlling for the extra knowledge gained of the pupils during the time between the two rounds of identification.

We have shown in Chapters Three and Four how differences that did exist between schools in the first round of identification showed no particular pattern that could be expected to impose bias in our random allocation to control and experimental schools. We will now give some attention to a detailed confirmation that no bias did, in fact exist.

Table 5.2 lists the schools that were allocated to the experimental and control groups.

TABLE 5.2

EXPERIMENTAL AND CONTROL SCHOOLS FOR PRELIMINARY STUDY

Experimental	01	02	03	04	10	11	12*
Control	00	05	06	07	08	09	

*10, 11, 12 were the three halls of school 13.

For the following comparisons we have pooled the results of the schools within the two groups. This was considered valid in the light of the results that have already been discussed in Chapters Three and Four concerning schools separately. These reults showed that differences between individual schools were small. After pooling the results of the schools in the two groups, in this way, we checked for any overall bias that may have occurred by chance if, in the random allocation, schools with small but consistent characteristics were placed in one group and schools with small but consistent characteristics of a different kind were placed in the other. The main gist of the results was that random allocation had been successful in producing two equivalent groups of schools.

The first analysis showed whether control and experimental schools differed, in Round 1, in the extent to which teacher-based and test-based assessments of pupils were in agreement. To carry out this analysis we counted up the number of pupils in the two groups of schools who were nominated for the top 10% group both by the teachers and by the tests. We also counted up the numbers who were nominated by the teachers but not by the tests. We then carried out a comparison between the numbers for the two groups of schools using the chi-squared test. This would show whether the two groups of schools differed in terms of test - teacher agreement. Also included in the table are figures showing the percentage of test-identified pupils who were also nominated by teachers.

TABLE 5.3

DIFFERENCES BETWEEN TEST AND TEACHER AGREEMENT
IN THE EXPERIMENTAL AND THE CONTROL SCHOOLS

Subject	X^2 (df = 1) (corrected for continuity)		% of test identified pupils also identified by teachers	
			Control	Experimental
ENGLISH	0.20	nsd	63	59
FRENCH	0.93	nsd	55	47
PHYSICS	0.01	nsd	44	44
MATHS	1.06	nsd	64	56

The results confirmed that the random allocation of schools to these two groups had achieved its end in producing two groups of schools which were very similar in the extent to which teacher and test judgements of pupils matched, based on this method of counting.

Table 5.4 shows the results of a similar analysis of the number of children in the high scoring A Group who were nominated by teachers in Round 1. There was no detectable difference when figures for the control schools were compared with those for the experimental schools.

TABLE 5.4

DIFFERENCES BETWEEN GROUPS OF SCHOOLS IN THE EXTENT TO
WHICH TEACHERS NOMINATED A GROUP PUPILS

	ENGLISH	FRENCH	PHYSICS	MATHS
		X^2 value		
Control schools against Expt Schools	0.47	0.00	1.59	0.44
Significance (df = 1)	nsd	nsd	nsd	nsd

We have discussed the importance of taking account of test error in previous chapters. One difficulty is, of course, that often very few marks separate people near the dividing line and therefore, in our context, false distinctions are made between nominating a pupil with a score of 120 who is above the line, and nominating a pupil with a score of 119 who happens to be below. Another difficulty is that all nominations 'above the line' are treated in the same way so that, in the work reported above, a teacher nominating the top 6 pupils out of a group of 12 above the line, would be judged to be equally accurate as a teacher nominating the lowest 6.

In an attempt to overcome these problems we have supplemented the above comparisons with an analysis which is based on the correlation between teacher judgement and predicted O-level score. No 'line' is referred to in this analysis which may therefore give a finer view of test/teacher agreement. These correlations were not reported in earlier chapters because the reader would have been overloaded with statistics all pointing to the same conclusions. Thus while we are making a particular point here regarding control and experimental schools, the reader may also wish to relate these results to those presented Chapter Three. For this reason full tables will be presented.

In Table 5.5 we give the values of the point-biserial correlation coefficients between teacher judgement and test-predicted O-level score. All pupils in the schools were included in this analysis.

TABLE 5.5

CORRELATIONS BETWEEN TEACHER JUDGEMENT AND TEST PREDICTED O LEVEL SCORE FOR ALL PUPILS IN EACH SCHOOL

SCHOOL	ENGLISH	FRENCH	PHYSICS	MATHS
00	.47 (145)	.49 (140)	.43 (149)	.52 (143)
01	.54 (205)	.44 (200)	.43 (205)	.51 (201)
02	.51 (170)	.43 (167)	.42 (173)	.52 (171)
03	.50 (138)	.39 (135)	.38 (135)	.42 (136)
04	.55 (174)	.29 (174)	.14 (173)*	.47 (172)
05	.46 (144)	.44 (143)	.46 (146)	.59 (147)
06	.58 (196)	.42 (196)	.43 (197)	.48 (197)
07	.49 (220)	.32 (222)	.28 (222)	.50 (218)
08	.50 (124)	.34 (122)	.30 (121)	.48 (122)
09	.60 (189)	.51 (190)	.51 (189)	.53 (194)
10	.60 (118)	.50 (118)	.47 (119)	.47 (116)
11	.49 (112)	.44 (110)	.31 (111)	.48 (111)
12	.46 (116)	.22 (116)**	.50 (118)	.51 (118)

* = sig at 5% level
** = sig at 1% level
all other correlations were significant at the 0.1% level.

All correlation coefficients in this table were statistically significant. Except where indicated this was at the 0.1% level. These significances in all schools in all subjects, provided evidence for a definite link between test and teacher judgements.

Since each school's third year can be treated as a random cluster sample from the same population (third year pupils in Oxfordshire) it is reasonable to find average correlation coefficients from the full set of school coefficients quoted in Table 5.4. This has been done by transforming the correlation coefficients into z scores, finding the mean z score and transforming back (Guilford and Fruchter, 1978). In Table 5.6 average coefficients for control and experimental schools are also reported separately.

TABLE 5.6

MEAN CORRELATION COEFFICIENTS

	ENGLISH	FRENCH	PHYSICS	MATHS
All Schools	.52	.41	.39	.50
Control Schools	.53	.41	.39	.50
Experimental Schools	.51	.40	.39	.50

Several points emerge from these tables in support of findings reported elsewhere. First the larger coefficients in mathematics and English indicated greater agreement between tests and teachers in these subjects than in physics and French. Secondly the range of coefficients in each subject indicated relatively small differences between schools except in physics. It is interesting to note that, in physics, in agreement with the result reported in Chapter Three, School 04 showed least agreement between test and teacher, and School 09 showed greatest agreement.

The similarity of coefficients in the control and experimental schools confirm the lack of difference between these schools that was reported above.

Finally we can look at the characteristics shown up as teacher biases within the two groups. For greater detail the reader should refer to the discussion of the method given in support of Table 4.12 where similar comparisons were made between groups of schools. Table 5.7 shows those few test scales which yielded significant differences when control and experimental schools were compared.

TABLE 5.7

SCORE DISTRIBUTIONS IN THE TEACHER SELECTED GROUPS
COMPARISONS BETWEEN GROUPS OF SCHOOLS

GROUPS COMPARED	Significantly different distributions in the teacher selected groups in:			
	ENGLISH	FRENCH	PHYSICS	MATHS
Control V Experimental schools	-	-	Attitude I $\chi^2 = 9.42$ df = 2, 5% Control schools have fewer low scoring pupils LU $\chi^2 = 10.55$ df = 2, 1% Control schools have more high LU pupils	Neatness $\chi^2 = 6.42$ df = 1, 5% Expt. schools have more high scoring pupils

In view of the large number of statistical comparisons it is not surprising that a small number are significant; it is more important to consider the large number of non significant results.

In the light of the findings reported here, we are confident that our random allocation procedure divided the schools into control and experimental groups that showed no meaningful differences, either qualitatively or quantitively, when judged on teacher nominations for the first round identification. We can therefore consider any differences that are demonstrated in the results of the second round of identification to be valid indications of the part that checklists played in altering the identification procedure.

Before presenting a detailed account of the results of the second round of identification, we give a brief account of the way in which the checklists that we gave to teachers were developed.

5.3 Construction of the checklists

In producing an ideal checklist one would seek to ensure that the items on the list were valid as indicators of high ability in the specific subject under consideration, and that these items were also relevant to what can be observed in the classroom. One would also bear in mind that the method that the teacher adopts regarding the use of the checklist may influence its effectiveness.

In connection with these issues, the first stage of our study of checklist usage was intended to simulate the situation that would exist if teachers were to follow the kinds of recommendations that are often made, for example by county working parties such as those to which we have made reference in Chapter One. This approach meant that we were essentially making an assessment of the effectiveness of checklists when used in the 'normal' way. This approach was followed up, especially in mathematics, in a second investigation which built upon these preliminary findings.

A search was therefore made of the relevant literature and discussions were held with teachers and, where appropriate, with other subject experts. No attempt at a large scale validity exercise was made at this stage in any subject. Reference to the recommendations included in the Schools Council booklet on identification (Clarke, 1981) will enable the reader to see how closely this procedure matches that which teachers are often urged to use. Because of the different starting points for each of the subject checklists we will now give brief separate explanations of their evolution. The checklists, themselves, are included in Appendix 5.

The physics checklist

It is first important to note that the physics checklist was designed as an aid to teachers in forming their overall judgement of pupils' abilities, not as an instrument that would, through some system of scoring, generate a list of 'top 10%' pupils in some automatic way. In view of the results reported in Table 5.1 it was clear that physics teachers placed consideraable emphasis on the use of their own records, and on the results of tests and examinations. We therefore sought to design a checklist that would provide data which was complementary to this kind of information already used by teachers. Teachers' records consist mainly of information on the quality of the written work done by the pupil. Examination and test results at third year level give, mainly, information on pupils ability to remember facts and relatively simple scientific concepts and to apply these to the solution of fairly standard types of problem. We therefore sought to produce a checklist that should help to broaden the charcteristics which teachers might look for in marking written work, and that might help to structure the teachers' subjective observation of pupils' scientific activity in physics lessons. The fact that such a checklist made relatively few references to pupils' abilities to do the standard kinds of physics problems should not be interpreted as a sign that we regarded such abilities as unimportant, but simply as an indication that these were the abilities which were already being assessed through teachers' present records and through test results.

The first task in designing a checklist to do the task we have been describing, was to generate a set of items describing the sorts of activity that an able pupil might engage in in a physics classroom. A considerable influence at this stage was the list of items used by Clarke and Cutland, P. (1977), in connection with the Cleveland Project. The items on this Cleveland checklist were suggested by a group of secondary school science teachers, and modified through 'much discussion' following pilot use in seven schools. A rather similar list of items[1] was produced by Harlen (1977), again after discussion with groups of teachers. Further suggestions for items for our checklist were gleaned from the science chapter of the HMI report (1977) on gifted children, from the work of Semper and Dutton (1980), and from publications arising out of the Science Teacher Education Project (Sutton and Haysom, 1974; Jenkins and Whitfield, 1974). We first produced a composite list of items drawing on the insights from these various sources. We then compared our composite list with the list of 'Science activity categories' set down by A.P.U. (1978). This A.P.U. list was intended to set out categories of scientific activity which:

> are recognisable by, and have validity for, those involved in science education. This list is designed to serve the needs of those who set the questions for the A.P.U. monitoring tests. Nevertheless, its ulimate justification is that it defines the main aspects of school science teaching.

A list devised for such a purpose, through a major initiative aimed at monitoring performance throughout the ability range, clearly provided a useful means of cross-referencing the ideas arising from work more directly concerned with the highly able in science. It was encouraging that most of the categories of the A.P.U. list were represented by one or more items in our own composite collection of items. Where we had other sorts of items they were allowed to stand; where there were omissions new items were produced to make good the gap. Discussion with physics colleagues in the Department of Educational Studies followed and a final list of items was produced. This final list can be found in Appendix 5.

One should, perhaps, re-emphasise that the individual items on the checklist were not subjected to a formal validation procedure by us. We also suspect that, those items which were taken from the sources which we have mentioned above, had not been validated in any strict sense by their originators. However it is important to note that the items had arisen out of extensive experience of school science and that they had been very carefully considered by a wide range of teachers and other professionals involved in science education. We would argue therefore that this extended and widely based discussion could be expected to lead to items that were entirely sensible. One could also argue that the value of the items as indicators of high ability in physics could be inferred from a consideration of the nature of physics itself (and of school physics in particular).

One final word on item validity concerns the massive investment of time that would have been needed to prove that each item was a valid indicator of high physics ability. Even with its resources, the A.P.U. science team did not attempt a formal validation of its own list of science activity categories. Indeed in the first full report on science in schools (Schofield et al, 1982) the authors state that they do not claim that their approach based on science activity categories:

> rests on any particular empirical or theoretical basis: in particular it is not claimed that the various categories and sub-categories are fully validated constructs.

We would suggest therefore that the procedure which we followed generated a reasonable set of items with acceptable face validity, and that the resulting checklist was similar to others being recommended for use. Since we were basing no claims for the overall effectiveness of the instrument on the validity of the individual items, but were intending to assess this overall effectiveness directly through our own research, we felt that the list of items generated in this way was satisfactory and moved on to consider how the checklist could be refined and how the teachers should be invited to record their observations.

The refinement was to seek, where possible, to help the teacher who was to use the list by indicating the kinds of areas of the physics syllabus in which we thought these abilities might most easily be observed. For example we suggested that the item asking teachers to assess pupil's ability to observe complex phenomena and select relevant from non relevant information, might be noted particularly easily when pupils were working with ripple tanks. This attempt to relate items to particular parts of the curriculum also served as a further check that the items did indeed describe behaviour which a teacher might reasonably expect to observe as pupils worked at their normal lessons.

The question of how teachers should record their observations involved a choice between two possible alternatives. The first, used by Renzulli et al (1971), involves the teacher in recording how often (occasionally, never, almost always) the pupil displays the particular behaviour described by the checklist item; the second, used by Clarke and Cutlaand (1977) replaces, as the checklist item, the single descriptor of behaviour which is indicative of high ability by a bipolar description of an ability related scale. The teacher is then asked to locate the pupil's behaviour on this scale.

Both systems have advantges and disadvantages. For example: the 'ability scale approach' tends to make the meaning of the checklist items more clear and certainly makes the checklist more useful to the teacher as a means of keeping records about pupils from the whole range of abilities in his or her class. (Many of these pupils might never quite produce the fairly sophisticated levels of behaviour which would allow the teacher to record anything of interest on the Renzulli scheme). This widened applicability of the checklist makes it more likely that, outside the reseach setting, teachers would find it a worthwhile aid, and would be prepared to devote sufficient time to completing it for all of their pupils. On the other hand a four or five point ability scale used with a mixed ability or broad banded class simply may not discriminate finely enough to help the teacher to select the top 10% with confidence.

We decided that, since no clearcut decision could be made, we would try one system for two subject checklists and the other system for the other two subjects. Since the clarity of the meaning of individual items was not a major problem in the physics list we chose to adopt the 'Renzulli style' for physics. This was also adopted in mathematics. French and English checklists were designed in the 'Clarke style'.

The English and French checklists

The same outline process in determining the design of the checklists was followed on both these subjects, so it is appropriate to describe that for both before proceeding to consider each subject in detail.

Examples of existing checklists were first consulted in order to determine a rough model for both format and items (a rough model at best, because existing checklists were predominantly concerned with pupils of high general ability). Then a sample of about a dozen experienced teachers including heads of departments, research officers in the appropriate subject areas, and tutors at the Department of Educational Studies was consulted for each subject to gain an initial impression of the kinds of items that might be expected to appear in checklists aimed at eliciting information about able pupils in those subjects.

Having achieved a broad consensus of opinion over the nature of these items (with noticeably more success in English than in French), the relevant literature was studied for research based evidence that would either clarify, or justify, or reject the criteria for identifying high ability implicit in them.

At that stage, a first attempt was made at drafting checklists in both subjects. Having done this, the same sample of teachers was again consulted and, from the comments thus elicited, the first working formulations of the checklists were derived.

The English checklists

There is a considerable amount of research based literature dealing with what represents ability, progress and attainment in English, as well as with how this might be assessed. Much of this was useful in judging the validity of items proposed by teachers, but none of the literature actually deals with checklists of the kind that concern the project.

Initial guidance came from the A.P.U. work on Language Performance, especially with items concerning reading and writing. Categories of activity concerned with the comprehension of passages were suggested such as i) overall impression, ii) selection of information from a wide range of text, (iii) detection of implied information (i.e. in a single passage), and these provided the basis for items 10(i) and 10(ii) (in the first checklist). The kinds of factors indicated as relevant to writing included aspects relating to the ability to adapt writing for different purposes, in dimensions such as i) narrative/descriptive or reflective/analytical, ii) first/second-hand subject-matter, iii) literary or functional. Many of these distinctions pointed to important factors, but were considered by many of the teachers consulted to require an analysis of the written work produced by pupils that would be beyond the scope of a concise tool of assessment such as these checklists were meant to be.

Wilkinson and Hanna (1980), point to the crucial aspect of reader awareness in writing, which represents a significant stage of development beyond the characteristics displayed by immature pupils, for who writing is predominantly i) narrative or descriptive, ii) affirmative, iii) literal and factual, iv) self-orientated. The ability to move beyond the 'egocentrical viewpoint' into writing that is explicit and projects to an audience, is shown to be an essential quality of increasing sophistication in writing. This point is supported in an article in 'Language for Learning' (1979) on the work of the Crediton Project. This article also clarifies a considerable number of other items to do with stylistic aspects of children's written work. For instance, 'structure/organisation' (relating in part to syntactical sophistication: 'There will be a development from simple to complex, but the furthest point of development will lie in the control of the appropriate structures in relation to the semantic needs' p.63); verbal competence, cohesion, appropriateness, and effectiveness.

Aspects of affective development in writing are also indicated by the Crediton Project work, includion notions of 'Self, Others and Addressees ('... the imaginative leap into the mind of another so that one grasps what terms have meaning for him is a kind of empathy which must characterise effective communication.' p. 64). Further guidance on aspects of writing was provided by reference to 'Assessing Compositions' (London Association for the Teaching of English, 1965), the work of David Holbrook (1981) Schools Council work on assessment, and 'Criteria and Assessment in English' by H.G. Earnshaw.

Most of these sources also provided much useful background to aspects of reading, notably Earnshaw:

> ...the main criterion in assessing Comprehension ability is the understanding of what is explicit in the passage: according to the type of passage this may be ability to follow the stages of an argument, the course of events, the parts of a description. The candidate should be able

> to see not only the details separately but the connections and links, the structure of the passage and its cummulative effect ... in the upper range of ability candidates should be able to make inference...

and, with regard to literature,

> A genuine response based on the reaction of the individual reader and authenticated by reference to the text is always welcome as the mark of a candidate distinctly above average.

These comments refer specifically to pupils at 16+, but might fairly be considered to pupils showing signs of high ability at 13+, especially as the project has accepted potential performance at O-level as a criterion for selection as highly able.

The oral aspect was by no means intended to be undervalued in the checklists, but it was felt that only a limited number of items should reasonably be included, in view of the fact that oral performance is particularly hard to assess. A broad impression judgement (in two parts) was finally chosen, to be prefaced by Douglas Barnes' comment, from 'Situated Speech Strategies (the monitoring of oracy)', that:

> ...someone whose oral abilities are high is someone whose speech matches both his own purposes and his perceptions of others' attitudes and understanding.

Having identified the areas of ability that could validly be referred to in the items of the checklist - including items not hitherto mentioned that were commonly found in general high ability checklists, as well as in the recommendations of the teachers that were consulted, such as sense of humour - the next task was to select a strictly limited number of items that were a) particularly informative and b) genuinely observable. It would easily have been possible to develop a checklist that was four times as long, but the general feeling was that the checking of fifteen separate items was as much as should reasonably be asked of teachers, and therefore a process of elimination was necessary to produce a final selection. The items that were selected were supported by the combination of backing in the literature and approval from the sample of teachers. Obviously, therefore, there are many omissions of aspects of ability in the subject that could justifiably be considered relevant: in the end, it was hoped to produce a checklist that pointed usefully to a significant selection of major areas that, although not claiming to be comprehensive, did provide a broad selection of positive indications of ability.

The French checklist

Initial input of potential items from experienced teachers of French was lower than from those provided by teachers of English, on the general grounds that pupils at 13+ had not yet developed far enough in the subject to be capable of exhibiting a wide vriety of indicators of ability. Factors such as high motivation/application, good memory, and selfconfidence expressed in a variety of terms and, by implication, in a variety of metaphors constitute the full extent of potential items proposed in the first stage.

Turning to the literature, certain potentially revealing items were discovered. First of all, the three categories referred to in the previous paragraph received wide research-based confirmation, in work done both with native-English learners of French, and with foreign language learning in general. Pimsleur (1964) and Oskarrson (1975) provide the chief sources of information. A further important source of guidance was a series of discussions with the OERG Research Officer working on the B.P. Modern Languages Project at the Department of Educational Studies, Miss A. Miller.

Pimsleur, p. 126, says that:

> ...motivation is one of the two main determiners of success in foreign language learning,

and, referring to self-confidence, he says, on p. 133:

> ...the person who does well in a foreign language is one who is willing to get up and talk before others....

Memory received no less support, but frequently in terms of extreme complexity. The concept was frequently treated with great precision, and the single term that teachers quite justifiably use refers to a whole battery of skills to do with, for instance, means of learning (eyeminded or earminded; cf. Pimsleur p 123), verbal discrimination, comprehension, and motivation itself. Even so, it was considered so essential by all teachers consulted that its presence in the form of the overall term was included in the checklist.

The other chief items provided by the work of both Pimsleur and

Oskarrson are

i) auditory ability ('According to this investigation, there does exist a 'talent' for learning foreign languages - that is, a special factor beyond intelligence and industriousness which accounts for how well an individual succeeds in a language course. Our evidence indicates that this special factor is auditory ability, which may be defined as the ability to receive and process information through the ear.' Pimsleur p. 135;

ii) flexibility ('... the ability to accept and use a new language seems to involve a degree of flexibility ... tolerance for ambiguity and frustration needed to get over that annoying period when one cannot express oneself easily in the new language.' Pimsleur p. 130. This factor, according to Oskarrson, 'Measures a person's attitude towards change, and his readiness and ability to adapt to new situations.' p.6);

iii) ear-minded vs. eye-minded learning (perhaps '... a connection between modality and foreign language study. Presumably, ear-minded individuals are more likely to enjoy studying a language, and to be successful at it.' p. 125);

iv) verbal fluency (Oskarrson p. 5) also referred to as native language fluency (Pimsleur p. 150).

Refinement of these items was provided in consultation (see previous note) and additional items were added concerning i) sound/symbol correspondence (which refers to the extent pronunciation of French is perturbed by the written form of the language) and ii) rule awareness.

Out of these combined sources, the first version of the checklist was developed. Further refinements were provided by the teacher sample, but it was only after actual use in schools that more extensive refinements and alterations were seen to be necessary.

The Mathematics checklist

We will consider here the origins of the first form of the mathematics checklist, and discuss, in more detail, the validation of the final form in a subsequent chapter.

In our search for valid checklist items for mathematics we have scrutinised the relevant literature and, as in the other subjects, endeavoured to include items that are both valid and observable in the classroom. There have been many research initiatives to identify the characteristics of able mathematicians in particular aspects of the subject. Mathematics involves problem solving of a particular kind and there has been considerable research both in problem solving of a general kind and problem solving in various branches of mathematics, (for example, Thorndike, 1922; Brownell, 1935; Newell, 1972). These researches are generally concerned with testing a model or strategy that is employed in problem solving. From these studies, however, characteristics of able problem solvers emerge. Kulm and Bussmann (1980) in evolving a rationale for their own 'Phase Ability Model' list many of the characteristics that previous researches have discovered. One notable contribution comes from the work of V.A. Krutetskii (1976). Characteristics that he has listed as valid indicators of mathematical ability are much the same as other researchers have discovered independently in a variety of studies.

Krutetskii's work has been accepted widely both by researchers and classroom practitioners. This is evidenced by its inclusion as a main reference for those concerned with the identification of able mathematicians in the Schools Council Booklet (Straker, 1981). It is also the main reference in the characteristics of able mathematics pupils in the widely accepted Cockcroft Report (1982). It represents the best starting point from mathematical problem solving research that is available for the compilation of a checklist for classroom use. Krutetskii's work grew from what was already known of the characteristics of able children from the international literature, what mathematics teachers considered the important qualities and what characterised mathematicians who had achieved eminence. He used as his model of the problem solving process one involving the three stages: gathering information, processing information and information retention. He identified three types of mathematical mind: The Analytic who thinks in verbal, logical terms, The Geometric who thinks in visual, pictorial terms, and the Harmonic who combines both of the other two characteristics.

By constructing ingenious challenges of particlar kinds and evaluating the responses of pupils of differing mathematical ability, Krutetskii was able to isolate the following characteristics.

1. Obtaining mathematical information
   - A. The ability for formalized perception of mathematical material, or grasping the formal structure of a problem.

2. Processing mathematical information
   - A. The ability for logical thought in the sphere of quantitative and spatial relationships, number and letter symbols; the ability to think in mathematical symbols.

   - B. The ability for rapid and broad generalizaton of mathematical objects, relations and operations.

   - C. The ability to curtail the process of mathematical reasoning and the system of corresponding operations; the ability to think in curtailed structures.

   - D. Flexibility of mental processes in mathematical activity.

   - E. Striving for clarity, simplicity, economy, and rationality of solutions.

   - F. The ability for rapid and free reconstruction of the direction of a mental process, switching from a direct to a reverse train of thought (reversibility of the mental process in mathematical reasoning).

3. Retaining mathematical information
   - A. Mathematical memory (generalized memory for mathematical relationships, type characteristics, schemes of arguments and proofs, methods of problem solving, and principles of approach).

4. General synthetic component
   - A. Mathematical cast of mind.

If it is accepted that these characteristics are valid they can form the basis of the checklist for teachers. It must be stressed that these are items referring to aptitude and it could be argued that attitude and motivation should be important components of the identification process. This will be considered elsewhere, but in the formation of the checklist, which is concerned with the identification of ability alone, leaving the exploitation of this ability through motivation as an important but secondary matter, only these items have been considered.

At this stage the main concern was that which was discussed earlier, of making the checklist into a form where the items related to what might be easily understood and observable by the teachers. This was achieved to some extent by looking closely at the sorts of challenge that Krutetskii gave and deriving classroom observable items related to them. The list often included, therefore, a number of items each of which related to the same mental ability from Krutetskii's list. For example items 5 and 6 of the checklist 'able to work and think fluently when symbols replace numbers' and 'able to transfer rapidly from concrete cases to generalisation' refer to Krutetskii's item 28, 'the ability for rapid and broad generalisations'. The teacher is much more able to observe one or both of the checklist items than to be able to relate directly to the single item on Krutetskii's list. Section E of the checklist, concerned with attitudes, was included so that the teacher's mind might be drawn to factors that could be distracting influences in his/her judgement.

5.4 Method of use

Appendix 5 includes the instructions that were sent out to teachers concerning use and completion of the checklists. The design of the procedure to be adopted in the use of these lists had to be a compromise between that which might ideally have been done and that which was practicable. Ideally we would have asked for checklists to be filled in on all but the very weak third year pupils. This would have ensured that a full analysis would have been possible of the ability of the lists, and of the individual items, to discriminate between the more able pupils and the rest. It would however have been quite unreasonable to expect this much work of the teachers. We therefore chose to ask for checklists on a smaller group of pupils roughly the top 20% as selected by the teachers. It was felt that this group should be large enough to draw in many of those pupils who did well on the tests but were omitted from the first round teacher lists and that it was small enough to represent a reasonable amount of work for the teachers. However the consequent loss of ability to assess discriminatory power of the checklists was disturbing, and we therefore added to this top 20% teacher chosen group a small number of named pupils chosen at random from pupils whose VN score on the Differential Aptitude Tests lay between the 45th and 55th percentile.

The other aspect of procedure to which much thought was given was the advice to teachers on how to use the checklists, once they had been completed, to narrow the original 20% group down to the 10% group which we were aiming for. Renzulli et al (1971) in their checklist adopted a scheme of assigning a numerical score to each column on the list and summing these scores to find the most able. Such an approach, while attractive in its simplicity, appears to overlook crucial things. First some items might not have equal weight compared with other items. Secondly some items may be far more relevant (or more accurately observed) in one school than another. Thirdly it seemed inappropriate to adopt such a rigid system in scoring an instrument which was intended to guide teachers' judgement (rather than to replace it). We therefore used the procedure described in Appendix 5 which essentially left the teacher to interpret the list in the way which seemed most appropriate in the light of the other information which was to hand. Again this procedure could only be thought of as tentative and was open to scrutiny once the lists were returned. We must, again, stress that we were monitoring the checklist use in the way that a teacher could normally have been expected to use them. We were, of course, careful, as will be shown later, to monitor the method of use so that the follow-up work could be most effective.

Two final comments will be made concerning the validity of the checklists before we go on to examine their effectiveness. First, we were careful to include items in all subjects that were relevant to the teaching aims for that subject; and secondly, the reader may wish to be reminded that it was the effectiveness of the overall instrument that was under scrutiny and not the effectiveness or validity of particular items.

Note for Chapter 5

1) It seems that these lists were not developed entirely independantly of one another.

CHAPTER SIX

The Effect of the First Version of the Checklists

6.1 Introduction

In this chapter we are concerned with two main issues. First, the effectiveness of the checklists which we described in the previous chapter, in improving teacher-based identification; and secondly the extent to which knowing pupils for a longer period of time enabled teachers to make more accurate judgements about which of those pupils had high potential in the subject. We have explored the first issue by comparing the results of teacher-based identification in the control schools with those achieved in the experimental schools during the second round (when the teachers in the experimental schools, but not the control schools made use of checklists). We have explored the second issue by comparing the results of the second round of identification with those of the first - which was carried out some four to five months earlier. In both sets of work we followed the same pattern of analysis as for Round 1 results. We therefore comment on test-teacher agreement; on the accuracy of teacher-based identification; and on the nature of some of the characteristics of the teacher-selected pupils in Round 2.

Before giving the detailed results we will summarise the main findings.

6.2 Summary of the results

a) The second round of identification showed little change compared with the first in terms of a broad measure of the agreement of teacher nomination and test nomination. This was true in all schools. There was some improvement concerning the numbers of highest scoring (A group) pupils who were selected by teachers. Apart from this effect, knowing pupils for longer did not seem to produce greatly improved teacher-based identification.

b) The levels of mismatch between test and teacher in Round 2 gave the same level of concern in physics and French as in Round 1. The agreement in mathematics and English in Round 2 was still at the encouragingly high level that we found in Round 1.

These results applied to both checklist, and non checklist aided identifications.

c) As suggested by the comment above, checklist usage did not seem to have a large effect on the match between teacher-based and test-based identification. The few differences that were detectable were all in favour of non-checklist identification. This was not a strong enough effect to enable us to say, with certainty, that non-checklist identifications were more accurate, but it does allow us to say quite firmly that use of the checklists described in Chapter Five, in the manner described there, did not lead to more accurate identifications.

d) When the quality of change of the top 10% lists from Round 1 to Round 2 was considered (e.g. in the sense of pupils selected by teachers in Round 1 despite a low test score being replaced, in Round 2, by higher scoring pupils) no differences were apparent between checklist and non checklist schools in any subject.

e) The percentage of pupils who were nominated by teachers in one round but not the other was high in all subjects. The very high change for physics was of concern. This was mainly because the change did not seem to lead to an increase in the number of 'good' selections (i.e. selections in agreement with pupils' test scores). In French, though overall there was still cause for concern in the number of weak pupils chosen, the changes that did occur indicated some improvement. In English and mathematics though overall accuracy did not change, there was some evidence of uncertainty concerning low scoring pupils as individual low scoring pupils were replaced, in the Round 2 lists, by other similarly low scoring pupils.

f) In terms of the characteristics that could be associated with the pupils who were identified, the same pattern was discernable in Round 2 as in Round 1. In both checklist and non-checklist schools, differences from teacher to teacher were evident and this suggested different biases on the part of individual teachers. The nature of most of these individual biases, which affected but did not completely determine teachers' judgements, suggested that there was some room for improvement in the teachers' identification strategies. There was, however, no dominant bias (such as a bias towards one sex or towards one social class) that could be identified consistently across all of the schools, and that might be of particular concern.

g) The method of checklist use that was employed seemed not to be effective.

6.3 Level of agreement between test prediction and teacher judgement

These comparisons were with the same rank ordered test predicted lists as for Round 1 with just the minor modifications made necessary by the small number of pupils who had left a school during the intervening months.

We have already stressed that in these simple measures of agreement between test and teacher prediction, we are not able to give confident conclusions about the accuracy of teacher judgement. It is worth remarking, however that the accuracy judgements that we were able to give in the later work from the Round 1 data did prove to be in line with inferences that could have been made from our study of agreement between test and teacher.

We will first consider the change of agreement between teachers and tests from Round 1 to Round 2. In Table 6.1, the results of all of the schools were combined. The table gives the percentages of those children who were identified by the tests, who were also identified by the teachers.

TABLE 6.1

PERCENTAGE OF TEST IDENTIFIED PUPILS ALSO IDENTIFIED BY TEACHERS

	ENGLISH	FRENCH	PHYSICS	MATHS
Round 2	61	54	45	61
Round 1	61	51	45	61

A fascinating aspect of this table is the lack of any real improvement between Round 1 and Round 2. It is, of course, possible that these overall figures mask an improvement for some schools and a worsening of the situation in others. Such a variation would be particularly interesting if it occurred between the experimental and control schools for conclusions could then be drawn about the effect that checklists had on the second round of identification. To investigate this possibility we repeated, on the second round data, the analysis reported in Section 3.3. Table 6.2 indicates the results of this work.

TABLE 6.2

DIFFERENCES BETWEEN TEST AND TEACHER AGREEMENT IN THE EXPERIMENTAL AND CONTROL SCHOOLS

Subject	ROUND TWO DATA			
	χ^2 (df = 1) (corrected for continuity)		% of test identified pupils also identified by teachers Control	Experimental
ENGLISH	0.09	nsd	63	60
FRENCH	0.01	nsd	55	53
PHYSICS	0.05	nsd	47	44
MATHS	1.85	nsd	66	55

There is clearly no evidence here that the agreement between test and teacher judgement was any greater, in Round 2, in those schools which had used checklists. The percentage figures quoted in Table 6.2 can be compared with similar figures in Table 5.3.. This enables us to get a first idea of the scale of the change between Round 1 and Round 2. It is tantalising that the agreement in the experimental schools was slightly better in Round 2 than in Round 1 (except in physics where it was unchanged) whereas in English and French in the control schools the match was unchanged between Round 2 and Round 1. It may be that there is some very slight advantage in the use of checklists in English and French but it must be emphasised that, at this stage, there is no firm evidence that this slight advantage is anything more than could be expected simply as a result of chance.

On the basis of this fairly crude measure of teacher/test agreement we are forced to conclude:

i) that knowing pupils for an additional term brought about no substantial improvement in the match between test and teacher judgement;

ii) that checklists did not lead to a sigificant difference between control and experimental schools in terms of this agreement.

In the style of the analysis in Chapter Three we also explored the Round 2 data to look for differences between individual schools, between 11+ entry and 13+ entry schools and between schools of different size. No significant differences were found.

Analysis of test/teacher agreement taking some account of test error

This work is based on a repeat of the analysis reported in Section 3.5. Table 6.3 shows the percentage of Round 2 teacher nominations falling into the test defined categories of 'good candidate for the 10% group', 'satisfactory candidate for the 10% group' and 'unlikely candidate for the 10% group'. Round 1 Figures are included in brackets for comparison.

TABLE 6.3

PERCENTAGE OF TEACHER NOMINATIONS FALLING IN THE THREE GROUPS 'GOOD', 'SATISFACTORY' AND 'UNLIKELY'

Subject	ROUND TWO DATA Percentage of teacher nominations who were: Good	Satisfactory	Unlikely
ENGLISH	61 (61)	25 (27)	14 (12)
FRENCH	54 (51)	15 (13)	31 (36)
PHYSICS	45 (45)	24 (25)	31 (30)
MATHS	61 (61)	25 (23)	14 (16)

It is interesting to note that there was no noticeable swing of teacher judgements away from the 'unlikely' candidates in Round 2. The only subject which showed the sort of changes whih might have been expected was French and even here the differences were small.

Having established this overall picture our next task was to determine whether control and experimental schools differed in the way in which teacher nominations were distributed across these categories.

Table 6.4 shows that no statistically significant differences could be established. As in the previous section we are led to conclude that neither longer aquaintance with pupils nor use of checklists gave rise to closer match between test and teacher judgement.

TABLE 6.4

DISTRIBUTION OF TEACHER NOMINATIONS ACROSS THE THREE CATEGORIES: DIFFERENCES BETWEEN CONTROL AND EXPERIMENTAL SCHOOLS

	ENGLISH	FRENCH	PHYSICS	MATHS
x^2	0.71	4.02	0.13	2.09
(df = 2)	nsd	nsd	nsd	nsd

A Group and C Group pupils: Round 2 nominations

We recall that the A group was composed of children who scored so high that there was little doubt that they would remain in the top 10% when they sat their O-level and the C group were low scorers who were very unlikely to become members of the top 10% on actual O-level performance.

A Group

The percentage of A Group pupils who were nominated by teachers in Round 2 are show below. It is interesting to note that in all subjects these figures were higher than those relating to Round 1.

TABLE 6.5

PERCENTAGE OF A GROUP PUPILS NOMINATED BY TEACHERS IN ROUND 2 (ROUND 1 FIGURES IN BRACKETS)

	ENGLISH	FRENCH	PHYSICS	MATHS
%	86.4 (80.8)	72.8 (61.4)	63.9 (53.7)	91.1 (76.2)

It is encouraging that teachers' ability to recognise these most able pupils on the test lists did improve between Round 1 and Round 2. In mathematics and English particularly the pupils from this group who remained unidentified by their teachers were perhaps just one or two in number in each school.

Having established this improvement we next explored to what extent it could be attributed to the use of checklists. In Table 6.6 we quote the results of chi-squared tests used to explore differences in the extent to which teachers nominated A Group pupils in the control and experimental schools.

TABLE 6.6

DIFFERENCES BETWEEN CONTROL AND EXPERIMENTAL SCHOOLS ROUND 2 TEACHER NOMINATION OF A GROUP PUPIL

	ENGLISH	FRENCH	PHYSICS	MATHS
X^2	.09	.11	.01	.06
(df= 1)	nsd	nsd	nsd	nsd

Again we are forced to conclude that the checklists did not increase teachers ability to nominate pupils from the A Group to a sufficient degree to result in a significant difference between the experimental and control schools.

We can only suppose that the improvement between Round 1 and Round 2 which was apparent from the results in Table 6.5 was a consequence of the fact that, by the time Round 2 nominations were made, teachers had at least six months experience of the pupils.

C Group

When the percentage of Round 1 and Round 2 nominations falling within the C Group were analysed very small overall improvements could be seen in all but English. (In English the situation was very slightly worse as more weak pupils were included in the teachers' groups - though one must stress that the change was very small).. Again no differences were apparent between control and experimental schools. These findings are summarised in Table 6.7.

TABLE 6.7

C GROUP NOMINATIONS IN ROUND 2

6.7a) PERCENTAGE OF TEACHER NOMINATIONS FALLING IN THE C GROUP. ROUND 2 DATA (Round 1 in brackets)

	ENGLISH	FRENCH	PHYSICS	MATHS
%	13.1 (12.2)	30.9 (35.7)	30.2 (30.6)	13.7 (15.8)

6.7b) CHI-SQUARED ANALYSIS OF DIFFERENCES BETWEEN CONTROL SCHOOLS AND EXPERIMENTAL SCHOOLS

	ENGLISH	FRENCH	PHYSICS	MATHS
χ^2	0.07	2.36	0.05	0.81
(df= 1)	nsd	nsd	nsd	nsd

It would appear that knowing pupils for longer helps teachers to rectify the sort of mismatch which leads to able pupils being overlooked, but not to do very much (except perhaps in French) to rectify the sort of error in which less able pupils are judged to be of high ability.

6.4 A comparison of results in checklist and non checklist schools in Round 2, and of the results of each group of school with the results of Round 1.

This analysis provides a detailed comparison of the match between teacher and test judgement in experimental (checklist) schools and in control schools. It also provides the first opportunity for a statistical comparison between the match in Round 1 and that in Round 2. (Previous comments on the comparison of Round 1 and Round 2 have been based on a straightforward comparison of percentages, without the advantage of a statistical assessment of how likely it was that differences, of the scale that were noted, would occur by chance alone). For the Round 1/Round 2 comparison, results from checklist and non-checklist schools were compared separately with the results of the whole sample of schools in Round 1. The analysis described below was similar to that of Section 3.7 where, for each school and subject, we counted how many pupils had scores which lay in bands, each half a standard error wide, from the mark which defined the 90th percentile on predicted score. Within each band we then counted how many were nominated by the teacher and how many were not.

We will first present the table of figures and then discuss the comparisons that were made.

TABLE 6.8

NOS. OF PUPILS NOMINATED BY TEACHERS WHOSE PREDICTED SCORE LAY IN EACH BAND OF WIDTH ½ A STANDARD ERROR BASED ON A DATUM OF THE 90th PERCENTILE

		BELOW 90th PERCENTILE				ABOVE 90th PERCENTILE		
	ENG/MATHS	1½-2½	1-1½	½-1	0-½	0-½	½-1	1-2½
	FR/PHYS	1½-3						
ENG	CHECKLIST	2	4	11	18	17	17	16
	NO CHECKLIST	0	5	13	21	24	28	15
	ROUND 1	3	8	27	45	46	45	37
FR	CHECKLIST	16	8	8	7	20	7	12
	NO CHECKLIST	12	5	13	15	17	16	23
	ROUND 1	35	20	27	18	33	27	46
PHYS	CHECKLIST	8	19	18	16	20	21	11
	NO CHECKLIST	5	11	25	16	23	11	16
	ROUND 1	18	24	38	39	44	29	19
MATHS	CHECKLIST	0	5	20	20	24	24	18
	NO CHECKLIST	0	4	12	21	28	20	23
	ROUND 1	6	13	29	38	59	39	27

In each cell of Table 6.8 there are three numbers: the total number of identifications by teachers in checklist schools, in Round 2, in that band; the total number in non-checklist schools, in Round 2, in that band; and the total number of identifications for all schools, in that band, from the first round of identification. This Round 1 total was the figure that was reported in Table 3.18.

Table 6.9 gives the numbers of pupils in each band who were not identified.

TABLE 6.9

NOS. OF PUPILS NOT NOMINATED BY TEACHERS WHOSE PREDICTED SCORE LAY IN EACH BAND OF WIDTH ½ A STANDARD ERROR BASED ON A DATUM OF THE 90th PERCENTILE

			BELOW 90th PERCENTILE				ABOVE 90th PERCENTILE		
		ENG/MATHS	1½-2½	1-1½	½-1	0-½	0-½	½-1	1-2½
		FR/PHYS	1½-3						
	CHECKLIST		272	128	89	56	25	8	2
ENG	NO CHECKLIST		320	127	108	64	31	5	3
	ROUND 1		524	310	229	134	56	19	8
	CHECKLIST		301	86	41	32	28	8	5
FR	NO CHECKLIST		283	78	50	43	29	6	9
	ROUND 1		533	140	89	71	51	26	23
	CHECKLIST		398	191	138	85	38	19	4
PHYS	NO CHECKLIST		406	171	145	83	35	14	8
	ROUND 1		811	376	281	167	77	26	16
	CHECKLIST		298	149	111	75	32	11	2
MATHS	NO CHECKLIST		293	137	97	72	27	8	2
	ROUND 1		614	277	213	121	54	23	9

We are now in a position to make comparisons from the figures in Tables 6.8 and 6.9. In all but two cells it is possible to use the chi-squared test of Goodness of Fit[1] to test for significant differences between the distribution of numbers. In this, we are making judgements about the totals across all schools.

As background to this, we recall that in Round 1, there was no significant difference between the control and experimental schools in any of the tests that we carried out concerning test/teacher mismatch, neither were there any meaningful differences from school to school. For this reason we have totalled across all schools in order to get the background distribution against which control and experimental schools can be checked in the Round 2 identification to see if there was a significant change between the two rounds.

In all, three comparisons were carried out: checklist schools against Round 1, non-checklist schools against Round 1 and checklist against non-checklist schools in Round 2. The first two comparisons were looking for any changes from Round 1 to Round 2 and the third looking for any difference between the checklist or non-checklist schools within Round 2 itself.

All the tests were one row chi-squares and an example will illustrate the method. In Tables 6.8 and 6.9, from the 0-½ S.D. cells below the 90th percentile for English, the following, table could be extracted.

TABLE 6.10

	NOMINATED BY TEACHER	NOT NOMINATED BY TEACHER
CHECKLIST	18	56
NO CHECKLIST	21	64

We can check to see if the checklist school numbers fit the same distribution as the non-checklist school numbers by compiling a new table whose rows add to the same total and whose second row numbers are in the ratio 21:64.

We then carry out a chi-squared test on the resulting numbers. Table 6.11 shows the procedure for the cell which was taken as an example above.

TABLE 6.11

	NOMINATED	NOT NOMINATED
OBSERVED	18	56
EXPECTED	18.28	55.72
	X^2 = .0058	DF = 1 nsd

This was done for each of the bands, for each subject, in the three ways described above.

From the 84 comparisons there were very few significant differences (at, or above the 5% level). Table 6.12 lists these.

TABLE 6.12

SIGNIFICANT DIFFERENCES FOR CHI-SQUARED TEST APPLIED TO CELLS OF TABLES 6.8 AND 6.9

SUBJECT	BAND	COMPARISON	X^2	SIG
ENGLISH	.5 to 1 above	CHECKLIST V CONTROL	5.52	5%
FRENCH	.5 to 1 above	CONTROL V ROUND 1	4.18	5%
PHYSICS	.5 to 1 above	CHECKLIST V CONTROL	5.14	5%
MATHS	1 to 2.5 above	CONTROL V ROUND 1	3.85	5%

This small number of significances can be considered of little consequence in view of the large number of comparisons that showed no significance, though it is interesting to note that in each of them the difference was such that the control schools tended to be more accurate than the checklist ones.

The indications are the same as we found in the less detailed study of test/teacher agreement described in Section 6.3. There was no change in the way in which teacher-selected pupils were distributed among the various bands of scores, in any subject, over the few months between the two rounds of identification. This was so in both the checklist and non-checklist schools. The direct comparison of checklist and non-checklist schools in Round 2 confirmed what could be inferred from these results: namely that there was no difference between the two groups of schools. Checklists had done nothing in these terms to add to what teachers could achieve without them.

We recall from the Round 1 results that, within the limits that we could make judgements of teachers' accuracy from our information on the match between teacher judgement and our test results (with their associated error), we could not detect inaccuracies in teachers' judgements in mathematics and English, though the level of mismatch between test and teacher was large enough to cause some concern about teachers' accuracy in physics and French. Since Round 1 and Round 2 have now been shown to be closely comparable in terms of the match between teacher and test judgements, it is reasonable to suppose that these same conclusions will apply to the second identification in all schools, whether guided by checklist usage or not.

6.5 An investigation concerning the test scores of pupils who were nominated in one round but not the other

It was possible for us to make a judgement of another kind on the identification data from the two rounds. So far we have been concerned with comparisons of overall numbers of pupils without seeking conclusions about the characteristics of those pupils who were on one list but not the other.

We could conjecture that, on the whole, most pupils who were identified first time would be identified later, but checklists could influence the teacher to refine his/her judgements on just a few pupils.

In order to investigate this possibility, we classified pupils, who were on one list but not the other in the two rounds of identification, as 'good', 'satisfactory' or 'unlikely' choices. This was done in the same way as was described in Chapter Three (see Table 3.9).

The process was, then, to compare pupils 'gained' in the checklist schools with those 'gained' in the control schools and then do the same comparison for the pupils who were 'lost' from the teachers' lists in Round 2.

Tables 6.13 and 6.14 contain the results which were subjected to the chi-squared test.

TABLE 6.13

COMPARISON OF CONTROL AND CHECKLIST SCHOOLS IN THE QUALITY OF PUPILS 'GAINED' IN THE SECOND ROUND OF IDENTIFICATION

	ENGLISH			FRENCH			PHYSICS			MATHS		
	GOOD	SAT	UNLIK	GOOD	SAT	UNLIK	GOOD	SAT	UNLIK	GOOD	SAT	UNLIK
CHECKLIST	4	6	4	9	2	7	18	6	22	11	6	7
CONTROL	8	9	4	11	6	2	15	11	19	13	10	6
	$X^2 = 0.56$			$X^2 = 4.95$			$X^2 = 1.95$			$X^2 = 0.78$		
	DF = 2 nsd.			DF = 2 nsd.			DF = 2 nsd.			DF = 2 nsd.		

TABLE 6.14

COMPARISON OF CONTROL AND CHECKLIST SCHOOLS IN THE QUALITY OF PUPILS 'LOST' FROM THE TEACHERS LISTS IN THE SECOND ROUND OF IDENTIFICATION

	ENGLISH			FRENCH			PHYSICS			MATHS		
	GOOD	SAT	UNLIK	GOOD	SAT	UNLIK	GOOD	SAT	UNLIK	GOOD	SAT	UNLIK
CHECKLIST	4	6	3	6	3	19	15	8	16	9	5	9
CONTROL	9	10	4	10	2	12	13	10	19	12	8	4
	$X^2 = 0.31$			$X^2 = 2.49$			$X^2 = 0.51$			$X^2 = 3.02$		
	DF = 2 nsd.			DF = 2 nsd.			DF = 2 nsd.			DF = 2 nsd.		

In all cases there was no significant difference in the way in which teachers in control and checklist schools changed their lists of top 10% pupils when we estimate the quality, in terms of test scores, of those pupils who were identified in one round but not in the other. This confirms indications from the previous sections. The figures in the tables are totals for all schools in the two groups. Numbers for schools separately would be too small for analysis; indeed some of the cells in the above tables contain rather small numbers.

It is interesting to study the extent to which teachers did change names on their 10% lists. We will consider the number of new pupils on the Round 2 lists - the 'gains'[2]. Table 6.15 gives, as a percentage of the total number of pupils on the lists, those who were included in Round 2 but not Round 1.

TABLE 6.15

OVERALL PERCENTAGE CHANGE OF PUPILS ON 10% LISTS BETWEEN ROUND 1 AND ROUND 2 (CONTROL AND EXPERIMENTAL SCHOOL TOTALS COMBINED)

ENGLISH	FRENCH	PHYSICS	MATHS
33%	42%	80%	47%

These percentages are high in all subjects, so that we can be confident that teachers, in general, did refine their judgements between the two rounds of identification, but combined with previous results, we see that the use of checklists made no overall impact on this refined judgement.

As a final step in this work we have compiled Table 6.16 which shows how for all schools combined the 'gains' from Round 1 to Round 2 compared with the 'losses'. The chi-squared test shows whether the number of gains in the three categories was significantly different from the number of losses in the same three categories. It therefore indicates whether the quality of the pupils gained was similar to the quality of those who were lost.

TABLE 6.16

COMPARISON OF NUMBERS OF PUPILS IN TEST SCORE CATEGORIES 'GAINS' COMPARED WITH 'LOSSES' BETWEEN ROUND 1 AND ROUND 2 (TOTALS FOR ALL SCHOOLS)

	ENGLISH			FRENCH			PHYSICS			MATHS		
	GOOD	SAT	UNLIK	GOOD	SAT	UNLIK	GOOD	SAT	UNLIK	GOOD	SAT	UNLIK
GAINS	12	15	8	20	8	9	33	17	41	24	16	13
LOSSES	13	16	7	16	5	31	28	18	35	21	13	13
	$X^2 = 0.12$ DF=2 nsd.			$X^2 = 11.02$ DF=2 p=.01			$X^2 = 0.33$ DF=2 nsd..			$X^2 = 0.15$ DF=2 nsd.		

Only in French was there a significant difference between the the gains and the losses in terms of whether they were good, satisfactory or weak candidates for membership of the top 10% group. In French more good candidates were gained, and more unlikely candidates were lost from the teachers' lists, indicating that, in the changes from Round 1 to Round 2 were entirely reasonable. In the other subjects the changes resulted in the replacement of some pupils by others who, in test terms, were of similar quality. We should perhaps remind the reader that this slight improvement in French was not enough to show up clearly when all of the pupils who were nominated in Round 2 were compared with the full list of Round 1 nominations.

If we now look back to Table 6.15 in conjuction with Table 6.16 and take account of the previous conclusions concerning accuracy, we can make further judgements for each of the subjects.

In physics, the very high percentage change (80%) is a cause for concern. We have presented evidence that indicates inaccuracies of judgement in both rounds of identification. The high level of change in the teachers lists produced as many 'good' candidates as before and also as many 'weak' ones. A considerable uncertainty on the part of physics teachers must be assumed.

In French the overall effect between the two rounds was to maintain sufficient inaccuracies in the lists of pupils to cause concern, but it is encouraging that a study of the changes that took place indicates that a significant number of weaker pupils were replaced by better ones, indicating that for some pupils better knowledge of them was gained over the months between the two identifications.

In mathematics and English, we have already reported that, within the terms of the test prediction, Rounds 1 and 2 were equally successful. Certainly there were 'unlikely' pupils on teachers' lists, but the numbers were small enough that we could not conclude that the teachers had definitely made errors. Test error allows us to predict that there will be a number of low test scorers who eventually score high at O-level, and in the case of English and mathematics, this 'allowed' number was similar to the number which actually appeared on the teachers' lists. However, the fact that 13 low scorers were replaced by a different 13 low scorers overall between the rounds, does suggest some uncertainty on the part of the teachers, and some error in either one round or the other. The fact that there was uncertainty is clear; the possibility of error in one or other round rests on the following argument. If the low scoring choices were correct in one round (and test error would suggest that they may well be), but they were replaced by other low scoring pupils in the other round, then some of the low scoring pupils who were nominated in one or other round must be errors - otherwise we would have to accept a larger number of low scoring nominations as correct than the test results would suggest is reasonable.

This reasoning is also valid for the other subjects, in particular for physics, where the number of changes in the Unlikely category is much higher.

Finally we repeat the important conclusion, that whatever changes were effected from Round 1 to Round 2 in this aspect of the work, there was no evidence to support the view that checklists made any impact on the process.

We now complete this work by considering some of the characteristics of the pupils who were nominated by their teachers in Round 2.

6.6 Test based characteristics of pupils chosen in Round 2

Previous sections of this chapter have dealt with quantitative judgements concerning the second round of identification. Some judgements concerning the quality of selection will now be made. The method used here is the same as that used in Chapter Four. The work provides clear evidence concerning the characteristics of the teacher-selected pupils in terms of the individual background test scales. These firmly established characteristics then allow us to make some tentative suggestions about the nature of the teachers' selection strategies, even though the component parts of those strategies which are brought to our attention by the analysis may be used, by the teachers, in an entirely unconscious way. For example a teacher's top 10% list in physics may contain pupils who were good at spelling. Good spelling would be then, a clearly established characteristic of the selected pupils. We can then make some suggestions about why this might be so. This could be because the characteristic of spelling was correlated with some aspect of scientific ability that the tests did not measure directly, or it could be that the teacher, consciously or unconsciously, regarded correctly spelled work as 'good' work and therefore classified the pupil as more able. The fact that our background testing was very extensive means that a large variety of attitudes and abilities can be monitored in this work, and the nature of some of the characteristics that emerged enabled us to make a tentative suggestion that inappropriate criteria were influencing some teachers in some subjects.

Table 6.17 shows large numbers of test scales that showed association with selection in each school and subject. For the correlation calculations, pupils were scored 1 if selected in Round 2 and 0 otherwise and comparisons were with the whole of the rest of the year group in the school. This can be compared with the similar table (Table 4.1) from Round 1. Table 6.18 summarises the total number of schools and subjects showing significant correlation with each test scale.

		CREATIVITY			ATTITUDES								DIFFERENTIAL APTITUDE TESTS											
	SCHOOL	FLU	FLEX	ORIG	A	B	C	D	E	F	G	H	VR	NA	AR	CL	CS	MR	SR	SP	LU	NEAT	SEX	SC
ENGLISH	00						•							•	•	•		•	•	•	•			•
	01	•	•	•	•	•	•	•	•		•	•	•	•	•	•	•	•	•	•	•	•	•	•
	02		•					•				•												
	03						•					•					•			•	•	•		
	04	•	•	•	•	•	•				•	•	•	•	•	•	•	•	•	•	•			•
	05						•				•	•	•		•			•		•				
	06	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•		•	•	•	•	•	•	•
	07	•		•													•							
	08	•	•	•			•	•			•		•	•	•	•	•	•	•	•	•	•		
	09				•							•	•						•	•	•			•
	10																							
	11												•		•		•					•		•
	12	•	•	•								•	•	•	•	•		•	•	•	•			•
FRENCH	00						•							•	•	•		•	•		•			•
	01	•	•	•	•	•	•	•					•	•	•	•	•	•	•	•	•	•	•	•
	02		•	•				•	•			•			•					•	•			
	03		•				•					•					•			•	•		•	
	04				•	•	•		•		•		•	•	•	•		•	•	•	•	•	•	•
	05						•				•	•	•	•	•			•		•				•
	06	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•		•	•	•	•			•
	07			•													•			•				•
	08					•	•	•	•		•	•	•	•	•	•	•			•	•	•		
	09				•							•	•						•	•	•			•
	10																							
	11																							•
	12	•	•	•							•		•	•	•	•		•	•	•	•	•		•
PHYSICS	00						•							•	•	•		•	•					•
	01	•	•	•	•	•	•				•	•	•	•	•	•	•	•	•	•	•			•
	02	•	•			•		•	•			•	•	•	•					•	•	•		
	03						•					•		•			•			•	•			
	04		•		•	•	•	•	•		•		•	•	•	•		•	•	•	•	•		•
	05						•				•	•	•		•			•						
	06	•	•	•	•	•	•	•		•	•	•	•	•	•	•		•	•	•	•			•
	07			•													•	•	•					
	08	•	•	•	•	•	•	•			•	•	•	•	•	•	•	•	•	•	•	•		
	09				•							•							•		•			•
	10																							
	11												•		•				•					•
	12	•	•	•									•	•	•	•		•	•	•	•			•
MATHEMATICS	00						•							•	•	•		•	•		•			•
	01	•	•	•	•	•	•	•		•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
	02	•	•																					
	03						•										•			•				
	04		•		•	•	•	•			•		•	•	•	•	•	•	•	•	•	•		•
	05						•				•	•	•	•	•			•						
	06	•	•	•	•	•	•	•		•	•	•	•	•	•	•		•	•	•	•			•
	07	•		•										•	•		•	•	•	•				•
	08		•	•		•	•	•			•		•	•	•	•	•	•	•	•	•	•		
	09				•							•	•		•				•		•			•
	10																							
	11												•		•		•		•		•			•
	12	•	•	•	•						•		•	•	•	•		•	•	•	•	•		•

NOTES: 1) The chi-square statistic was used to determine significant correlations between teacher selection and sex, neatness and social class. The significance of the correlation coefficient was used for all other scales. Variables are ticked if the level of significance is 5% or higher (i.e. 1%, 0.1% etc.).

2) Correlations for sex always indicated that teachers tended to select girls.

TABLE 6.18

NO. OF SCHOOLS SHOWING SIGNIFICANT CORRELATION BETWEEN SELECTION AS 'MORE ABLE' BY TEACHER AND EACH BACKGROUND TEST (ROUND 2)

	ENGLISH	FRENCH	PHYSICS	MATHS
CREATIVITY				
FLU	6	3	5	5
FLEX	5	5	6	6
ORIG	6	5	5	5
ATTITUDES				
A	4	4	5	5
B	3	4	5	4
C	7	7	7	7
D	4	4	4	4
E	2	4	2	0
F	1	1	1	2
G	5	5	5	6
I	8	6	7	4
DAT				
VR	8	7	8	8
NA	6	7	8	8
VN	8	8	9	10
AR	6	6	6	6
CSA	6	4	4	6
MR	7	6	8	8
SR	7	6	9	9
SP	9	10	7	7
LU	8	9	8	8
SEX	2	2	0	1
NEATNESS(NT)	5	4	3	4
SOCIAL CLASS(SC)	7	9	7	8

1. Each section of school 13 was analysed separately. The maximum number of schools which could show significant correlations on any variable was therefore 13.
2. Only 11 schools were mixed. The total number of schools analysed with respect to sex was therefore 11. All correlations with sex were such as to indicate that teacher selected groups were biased towards girls.
3. All correlations with social class were negative indicating a teachers group bias towards the professional end of the social spectrum.

If we recall that the 'checklist' schools were (01, 02, 03, 04 and 13 -10, 11 & 12 being the separate units of school 13), we see from Table 6.17 that certain schools in both 'checklist' and control groups revealed interesting results - e.g. School 10 showed no significant correlation at all, in any subject, - but that there was no overall pattern to lead us to conclusions about effectiveness of checklists. All DAT scales except CSA featured strongly as did some attitude scales and social class.

It is important to follow this with an analysis that identifies the strongest characteristics as we did for Round 1. This was done by the same technique of Partial Correlation, in which we were able to identify the variables that appeared in the list only as a consequence of intercorrelation with some other variables and thus we could eliminate them as being secondary characteristics.

We did not consider it necessary, in view of the results already reported in this chapter, to repeat all of the work that we did for Round 1. Our main intention was to find clues to the inappropriate biases that might exist in the identification process. As we reasoned in Chapter Four, this could be best done by looking within Group B (the group of pupils who all scored high enough on the tests to indicate that they were plausible candidates for the top 10% group).

Table 6.19 gives the summary of characteristics.

TABLE 6.19

MAIN VARIABLES CHARACTERISING SELECTION FROM B GROUP AS SHOWN BY PARTIAL CORRELATION (PEARSON PRODUCT MOMENT CORRELATION IN BRACKETS)

SCHOOL	ENGLISH	FRENCH	PHYSICS	MATHS
00	CSA(0.59)	VN(0.68)	FLU(0.15) ORIG(0.08) B(0.34) VR(0.43)	NONE (AR 0.37)
01	NONE	NONE (FLEX 0.58)	NONE (B 0.22)	A(0.42)
02	----	NONE (E 0.37)	B(0.49) NA(0.44) C(0.41)	C(0.47) VN(0.39)
03	NONE (FLU 0.53)	NONE (A 0.60)	SC(0.45) SP(0.57)	A(0.55)
04	NONE (LU 0.41)	SP(0.31) LU(0.40)	B(0.21)	C(0.06)
05	NONE (MR 0.34)	NONE (E -0.66)	NONE (SP 0.58)	NA(0.62)
06	NONE (LU 0.46)	NONE (B 0.56)	B(0.51)	C(0.37)
07	NONE (SR 0.32)	C(0.45)	NA(0.33) C(0.22)	C(0.33) NA(0.37)
08	CSA(0.20)	NONE (NEAT 0.36)	NONE (FLEX 0.54)	NONE (CSA 0.60)
09	NONE (LU 0.40)	CSA(-0.33)	LU(0.42) VR(0.45)	I(0.27) NA(0.21)
10	AR(0.61) VN(0.53)	A(0.51)	F(-0.36) MR(0.54)	NONE (E 0.54)
11	ORIG(0.60)	----	NONE (SR 0.61)	A(0.16) E(0.47)
12	NEAT(-0.38)	SP(0.57)	I(0.21)	NONE (NEAT 0.45)

1. Sex being a categorical variable is not really suitable for this form of analysis and is omitted from this table.

2. If no variable could be isolated, the variable given, in brackets, is the one giving the highest product moment coefficient with teacher selection.

Table 6.19 should be compared with Table 4.15 in order to assess the change from Round 1 to Round 2. In all four subjects in Round 2 there were more schools in which strong characteristics could not be isolated. (7 against 3 in English, 6 against 2 in French, 4 against 0 in physics, 4 against 0 in mathematics.) Since we have argued that B Group analysis tends to bring inappropriate characteristics to light, this change can be interpreted as some improvement from Round 1 to Round 2.

Before giving a summary of the implications of these results it is useful to add one other set of figures. These concern the nature of the characteristics established by the comparison of the teacher-selected pupils with the whole of the rest of the year group (Table 4.2 for Round 1, and Table 6.18 for Round 2). We argued in Chapter Four that variables which did not show up in these comparisons as characteristics of teacher-selected pupils could, with some certainty, be regarded as characteristics which did not affect teachers' judgements. In Round 1 we found that, in several schools, the test scale that was most appropriate as a predictor for O-level in each subject failed to show up, and we commented that this indicated an undesirable feature of the teachers' identification strategy. In Table 6.21 we show the number of schools in Round 1 and Round 2 in which the most appropriate predictor failed to show up.

TABLE 6.21

NO. OF SCHOOLS WHOSE SELECTED GROUP WERE NOT STRONGLY CHARACTERISED BY THE BEST TEST PREDICTOR IN ROUND 2 (ROUND 1 IN BRACKETS)

ENGLISH (LU)	FRENCH (LU)	PHYSICS (MR)	MATHS (MA)
5 (4)	4 (4)	5 (9)	5 (5)

There seemes to have been little change in most subjects, but in physics fewer schools seemed to overlook the best physics predictor (MR) in Round 2.

Despite these improvements, the picture in Round 2 was still not entirely a satisfactory one. The strong characteristics that could be detected still varied from teacher to teacher as in Round 1, and there was still no detectable pattern in either checklist or control schools, or any noticeable difference between the two groups of schools - which adds to our other evidence concerning the ineffectiveness of the checklists when used in the normal way. The majority of strong characteristics (Table 6.19) were still not the most appropriate for the subject, indicating that a possibly inappropriate and illogical bias in the selection procedure had survived into the second round.

We can begin to evolve a theory concerning these changes from Round 1 to Round 2. In Round 1 the teachers' lists were characterised in an assortment of ways, some appropriate some not and we had the impression of individual teacher bias, much of which was detected by the background tests. There was some cause for concern in English and mathematics, but in many schools the strong characteristics were appropriate. This supported the view that there was reasonable agreement between test and teacher in these subjects. In French and physics, test and teacher agreement was less strong in Round 1 and a study of what characterised teacher identification revealed a larger number of inappropriate criteria than the other two subjects, particularly in physics where a strong bias towards the general ability measure VN was detected.

In the second round, in both checklist and control schools we detected changes in membership of the teachers' lists, though this led to little more agreement with the test prediction than had been discovered in Round 1. We also detected changes in the characteristics of the choices. Many of the new biases cannot be detected by our background test scales. We could suppose that over the intervening months the teachers have come to know the children in a different way and it is other qualities that caught their eye, but in the case of physics, for example, these newly observed characteristics built in bias that led to no better overall success of the identification than the Round 1 VN bias. There was no clear evidence that, used in the way in which they had been used, the checklists had any significant effect on teacher-based identification.

Notes for Chapter Six

1) The fact that the same children appeared in the Round 1 and Round 2 figures was a major factor in determining the kind of analysis that was employed in this chapter.

2) The total numbers 'gained' and 'lost' are a little different, for several reasons: Some children left a school between Rounds 1 and 2, sometimes pupils with missing data were involved in the list changes and sometimes teachers nominated one or two more in one round than another (rounding up or down for their total of 10%). These differences have no effect upon our conclusions, however.

CHAPTER SEVEN

The Interview Study of Pupils About Whom Tests and Teachers Disagreed

7.1 Introduction

On numerous occasions in the report we have, both in the nature of our analysis and in our interpretation of results, stressed that errors associated with tests have to be taken into account. We have made every effort to derive sound predictors for O-level, that were subject specific and were combinations of the most appropriate tests from a wide battery. Despite this, there are bound to be differences between predicted O-level scores based on test results achieved by pupils at 3rd form level and actual O-level scores achieved at 5th form level. Indeed this must, to some extent always be so whatever tests were used.

The fact that we had an estimate of average test error enabled us to compare the actual mismatch between test and teacher judgement with the mismatch that might be expected on the basis of the test error alone. In mathematics and English the actual mismatch was quantitively at the expected level. In physics and French there was more mismatch than could be accounted for simply by the error in the test. The implication is that, in these subjects especially, teachers also made some errors.

Added to this, in all subjects, the dominant characteristics of the teacher-selected pupils that could be detected from background tests were, in most cases, not entirely what we might expect if teachers were pursuing a sound identification strategy. Therefore, even in English and mathematics, there was some evidence that, although the number of low scoring pupils who were selected by teachers was rougly comparable to the number which could be regarded as reasonable in the light of the test error, the individual pupils who were chosen were pupils with perhaps rather unexpected characteristics. This suggested that the low scoring pupils whom teachers picked may not have been the low scoring pupils who would actually perform well at O-level - in other words that the teachers may have picked the wrong ones because their judgement was affected by inappropriate pupil charcteristics. A similar argument can be constructed in the case of high scoring pupils who were overlooked. This, of course, implies that the fullest possible study of the characteristics that influence teachers was needed.

We have, already, the insights gained in Chapter Four. However, despite these test based insights we could not be certain whether the teachers were also noticing characteristics in their pupils that the tests did not measure. Of course these additional characteristics that influenced teachers may or may not be appropriate to sound strategies. The accumulated force of these points is as follows.

In all subjects, we are left with questions concerning those aspects of the identification strategy that the test scores could not answer. In particular there were questions about those pupils who were extreme 'mismatches' in terms of teacher and test agreement. Since we suspect some degree of error on the part of teachers in all subjects (and especially in French and physics), the nature of these further aspects is of considerable interest. They clearly warrant further exploration - exploration that would necessarily take us away from any further study of the identification process in terms of the test results.

In order to make some headway we decided to seek the cooperation of teachers in discussing reasons behind their identifications; in particular reasons behind the decisions which led to the extreme mismatches involving pupils from Groups A and C. At the outset we were looking for explanations of these mismatches in an open minded way. The procedure which we used was to ask the teacher to talk about specific pupils. We included many Group A and Group C pupils about whom the teacher's opinion was at odds with the test result. As far as possible we tried, at this stage, not to reveal the test results for the pupils. In most cases there was therefore no question of the teacher feeling that he/she needed to justify his/her decision in the light of contradictory test findings (though very occasionally this aspect slips into the interviews - as will be seen in the following tables). We feel that the teachers were at ease to talk freely about the pupils and reasons for their choices, and that some interesting results emerged about the things that teachers noticed about pupils and the way in which these things contributed to their overall judgement of the pupils' potential.

7.2 Outcome

It was appropriate to use the second round identification lists to identify the appropriate mismatch pupils for this work. This gave the opportunity to look for extra clues regarding the use of checklists and also made use of the most recent opinions of the teachers in all cases.

We started with the intention of filling our gaps of knowledge, but what we found moved us to a more structured study. We might have expected the teachers to give wide descriptions of each child and perhaps so wide that we would find it difficult to assess which qualities were appropriate to the subject under consideration. For example, we might have heard about a choice for the mathematics list: 'Very good at geometry, vivid imagination, captain of chess club, plays piano at grade 6, always keen to appear in school play, enjoys the challenge of puzzles, scientific background at home ...' Instead of this we received very limited descriptions of the pupils. We were certainly given reasons for choices for and exclusions from the 10% lists, but at a very superficial level.

In all subjects, including those cases where checklists had been used where we might have expected much more subject specific detail, we were given just brief descriptions containing a limited number of separate ideas about each pupil. There was a bias towards generalities concerning both ability in the subject and attitudes and behaviour.

There were two possible reasons for the superficial nature of the responses. First, it may have been difficult for the teachers, in an interview situation, to verbalise all that they knew and felt about the pupils. Second, the knowledge of the pupils may have been limited.

We believed that both of these were true to some extent and we will describe, in succeeding chapters how we proceeded with our investigations to increase our understanding of the situation. There was, however, almost certainly, lack of the detailed subject specific knowledge that we had hoped for. If this was not so we would have had, within the descriptions of the pupils, much more detail. Rather than such generalities as 'good exam performance', we would get a series of descriptions such as, 'spatially aware', 'Good Mental Arithmetic', 'Never gives up with a problem', and many more such as these.

We now present a brief summary of the results of the interviews for each subject. We will not attempt to analyse them statistically. This would be inappropriate, particularly in view of the fact that this method of eliciting information may have been restrictive. We are, however, able to look at the tables and draw broad conclusions.

Because of the limited range of descriptions that emerged in the interviews, we have attempted to group the qualities that the teachers mentioned into the following categories: specific subject abilities; general intelligence; attitude and behaviour.

For the purposes of brevity we have used the following codes to describe the individual pupils:

M - Miss. Group A pupil not on teacher's list.
H - Hit. Group A or B pupil on teacher's list.
FH - False Hit. Group C pupil on teacher's list.

The two types of most interest, of course, are M and FH.

We will discuss more fully, in the next chapter, how we began to consider these interviews with teachers as a preliminary to a more detailed exploration of the identification strategies in the second year of the research.

7.3 Results for English

Table 7.1 gives a summary of the interviews with English teachers. Each section of the table gives the teacher's comments about one pupil. The pupil's 'type' (Hit, Miss, or False Hit) is indicated in the left hand column.

TABLE 7.1

DESCRIPTIONS OF PUPILS IN ENGLISH BY THEIR TEACHERS

PUPIL TYPE	DESCRIPTION	ENGLISH SKILLS	GENERAL ABILITY	ATTITUDE BEHAVIOUR	OTHER
School 03 (Checklist)					
H	Quite bright orally. Shows off a bit. Mind runs ahead of his hand. Very ordinary - but there is the odd flash.	1	2	1	
FH	High sensitivity to English.	1			
M	Conscientious. Creatively thoughtful.		1	1	
M	Good sense of humour. Tends to be idle. Doesn't perform as well as capable.		1	1	1
	Neat type. No real spark		1	1	
School 09 (Not Checklist)					
H	Not achieving in English. Extremely bright. Writes plays. Keen on computers. Won't do homework.	1	2	1	
M	Extremely bright. Fluctuating at first. Now consistent.		2		
FH	Orally V.good. Written work careless. Rather immature. Supportive home background.	2		1	1
	Weakness in skills. Doesn't read well. V.weak but has improved. Finds it difficult to put thoughts into writing.	3			
M	Quiet. Not entirely accurate		1	1	
	Fitfull skills. Not outstanding.	1	1		
	Very articulate. Hopeless speller. Poor skills - very bright.	2	1		

Table 7.1 is continued.

Table 7.1 Continued.

PUPIL TYPE	DESCRIPTION	ENGLISH SKILLS	GENERAL ABILITY	ATTITUDE BEHAVIOUR	OTHE
School 13 (Checklist)					
FH	Does well without hard work. Natural ability but lazy. No interest in literature. Likes what's easy. Accurate on technical skills.	2	2	2	
M	Quite hardworking. Bright and humorous. Useful boy to have in classroom.		1	1	2
M	Careless at first. Now trying hard. Tremendous reader. Highly independent. Trying to overcome punctuation problems. Spelling not too bad.	3		2	1
M	Uncommunicative. Lacks concentration. Mechanically accurate. Writes immaculately.	2		2	
M	Immature.			1	1
M	Keen. Not outstandingly bright. Articulate. Normal for age.		3	1	
M	Bright. Superficial interest in English.		1	1	
School 07 (Not a checklist school)					
FH	O level girl. She is good.		1		1
FH	Outstanding. A level material.		1		1
M	Not particularly good at English. Just in O level group.		1		1
M	Poor old Stephen! Dubious 'O' level. Not a shining light. Orally good. Poor control over writing.	1	1		1
M	O level candidate. Quick, perceptive, sensitive. Smashing kid. Enjoys poetry. V. poor home background.	1	2	2	2

TABLE 7.2

SUMMARY FOR ENGLISH
Number of comments in Each Category

	ENGLISH 'SKILLS'	GENERAL ABILITY	ATTITUDE/ BEHAVIOUR	OTHERS
FH	8	4	3	3
M	11	20	14	7
H	2	4	2	-

Conclusions:

These results suggest that teachers are taking greater notice of their impression of English 'skills' than of the other criteria when they select (FH) pupils (i.e. when they make the judgement that pupils are brighter than the tests indicate); whereas they either ignore an impression of general ability, or they consider that the pupil's perceived general ability does not transfer to English ability, when failing to assess pupils as possessing the ability that the tests indicate (Miss pupils). The number of English 'skills' mentioned for any child, however, is very low.

Judgements about attitude/behaviour relate more or less evenly to both False Hits, and Misses, and are certainly present to effect judgements that are made. A difference between the types of mismatch could be concerning the relation of impressions about English Skills and General Ability to a pupil's potential. It does appear that teachers are more inclined to make a positive mis-identification (assuming the accuracy of the tests) when considering specific English 'skills', and a negative mis-identification when considering General Ability.

It is, however, possible that, in a few cases, this information could be perverted by the fact that some teachers were possibly attempting to justify their mis-identifications by talking in these terms: by acknowledging General Ability in Misses and suggesting a lack of English skill nonetheless; and by insisting on some perhaps spurious English skill outside overall ability when referring to False Hits. Either way, the division in the impression created by the two separate criteria of general and specific ability is interesting, because in English at any rate, it seems like a false division.

7.4 Results for French

TABLE 7.3

DESCRIPTIONS OF PUPILS IN FRENCH BY THEIR TEACHERS

PUPIL TYPE	DESCRIPTION	FRENCH SKILLS		PERCEPTIONS ABOUT GENERAL INTELLIGENCE	ATTITUDE/ EFFORT/ BEHAVIOUR	OTHER
		TECHNICAL ACCURACY	ORAL CONFIDENCE			
School 04 (Check-list)						
H	Mature, confident, conscientious over detail. Orally strong.	1	1		2	
M	Gives the impression in class of not being interested. Fairly successful in written work.	1			1	
M	Mature. Lack of application, inclined to make errors. Orally confident.	1	1		2	
M	Careful, hard-working, very shy - orally weak.	1			3	
M	Lack of attention to detail, inclined to make mistakes in written French. Orally stronger.	1	1		1	
FH	Shows enthusiasm for the subject, takes care - not highly intelligent.	1		1	1	
FH	Quiet, sensible girl. Very willing. Orally weak but accurate in written work.	1	1		2	
FH	Apparently hard-working but occasionally inclined to make errors of grammar.	1			1	
School 06 (Not checklist)						
M	Fast thinker. Home					

Table 7.3 is continued

	problems. Badly inaccurate in writing.	1		1		
M	Reasonably careful, but shows no real aptitude for using a foreign language.	1			1	
M	Speaks faster than he thinks. Over-confident.		1		1	
H	Highly able. Orally brave and alert to details of written French	1	1	1		
FH	Orally impressive enthusiastic member of class. Little understanding of grammar.	1	1		1	
FH	Good accent. Little care in thinking.		1		1	
FH	Highly conscientious. Not very clever but makes an effort in French.			1	2	
H	High concentration. Imaginative use of French.	1			1	
School 09 (Not checklist)						
M	Very intelligent boy, but with big problems. Low level of commitment, careless, little inclination to work. Capable of succeeding in French.	1		1	2	1
M	Superficially able. Over-confident, pays little attention to detail.	1		1	1	
M	Bright but shows little understanding of the requirements of French, not serious.	1		1	1	
FH	A steady worker: easily capable of O level, careful and accurate. Orally fair.	1	1		1	1

Table 7.3 is continued

FH	Confident and enthusiastic, well able to cope with grammar.	1			1	
School 03 (Checklist)						
M	No real prospect in French. Chronically inaccurate. Orally weak.	1	1			
M	Orally fair. Written work sloppy.	1	1			
M	Has become lazy since beginning of year. Not committed.				2	
H	Enjoys French. Conscientious girl. Very accurate, technically.	1			2	
FH	Able. Successful in tests.			1		1
FH	Orally strong. Enjoys the subject. Potential 'O' level.		1		1	1
FH	Reliable worker. Not outstandingly clever, but very well motivated.			1	2	

TABLE 7.4

SUMMARY FOR FRENCH
Number of comments in each category

	FRENCH 'SKILLS'					
	Technical accuracy	Oral Confidence	TOTAL	Perceptions about general intelligence	Attitude/ effort/ behaviour	Other
FH	7	5	12	2	13	2
M	11	5	16	4	15	2
H	4	2	6	1	5	-

Conclusions:

Although it is initially tempting to consider the two aspects of *French skills* (technical accuracy; oral confidence) as two halves of one unit, the terms that French teachers use to discuss their pupils suggest that comments about *oral confidence* relate more closely to comments about attitude/effort/behaviour (with the exception of when *accent* is praised specifically). In the majority of cases, comments on oral confidence describe the pupil's *willingness* rather than *skill*.

Thus - the only generalisation that can be drawn from these limited figures is that in *all* cases included here, the majority of teacher's judgements (in all three categories, close to 2:1) are made on the basis of the pupil's apparent confidence/enthusiasm, rather than on the basis of the pupil's technical accuracy. Nonetheless, technical accuracy still figures highly, and other criteria, such as perceptions about general intelligence, problems at home etc., do not figure significantly.

As in English, however, the number of different subject specific descriptions of their pupils was much lower than would have been expected from detailed knowledge of the pupils.

The conclusions about confidence/enthusiasm are in accord with that found from the statistical work reported earlier.

7.5 Results for Physics

TABLE 7.5

DESCRIPTIONS OF PUPILS IN PHYSICS BY THEIR TEACHERS

PUPIL TYPE	DESCRIPTION
School 00 (Not a checklist school)	
FH	Lazy worker. Slipping now, not committed, very casual. Father pushes. Showed interest and put in a lot of work on an open evening. Backs away from school work.
FH	Was picked as a result of exam, other information would not confirm. Not standing out in Yr4 2nd set. Doesn't ask/answer much.
FH	Not a stroke of work in year - did well in exam. Not outstanding in Yr4, not doing homework, smoking generally getting into trouble. Brother in Oxford and suffers from comparison. Chairman of Governors daughter.
FH	Went off to end of 3rd Yr. Now reckon him to be top 20 not top 10. Enthusiastic, hard working, seems to have aptitude but struggles with subject. +ve attitude.
FH	Intense concentration, very good work, enthusiastic, making good use of what he's got. Not doing quite as well now.
M	Dramatic improvement this year, not industrious. Responds well to mathematics side, quiet, overshadowed by dominant brother and sister.
M	Good trier, very quiet, better exam than classwork. Mathematics now bringing her down.
School 01 (Checklist)	
FH	Crashed during year, can't cope with mathematics. Neat, tidy. 'Got to be impressed by neat orderly work, its half the battle.' Good on early tests.
FH	Pulled up by twin. Did well in Christmas test but fell to 24th in class later.
FH	Up from lower form ½ way through year and has pulled up well. Has some ability - caused a lot of problems because of attitude. Not doing much in Yr4.
FH	Good at start, went of in exam - 'girl trouble' - with not much will to exert himself.
FH	Very neat, gives impression of higher ability than she's got.

Table 7.5 is continued.

Table 7.5 continued.

PUPIL TYPE	DESCRIPTION
School 01 (continued)	
(Checklist)	
M	Peculiar attitude, cheeky, flippant, now in 'real thing' and getting on.
M	No impression made, soaked it up, no question in class didn't stand out.
M	Silly, didn't work hard, come on a lot this year. Excellent general knowledge.
M	Things a 'bit of a joke' - asked plenty of questions.
M	Life was a laugh - been in various sorts of trouble, now way out in front.
School 03	
(Checklist)	
FH	May need help to get a concept straight at first but can act on it and apply it soundly once its been made clear.
FH	Asks pertinent questions, same as previous girl
FH	'Intelligent' but not mechanically minded - A good sponge without solid foundational knowledge or appreciation. Would get by in physics through retention rather than understanding.
FH	Strong on mathematics, good at prac. Has +ve attitude (homework in on time). Gives impression of being there to find out what its about. Interested in applic. of Physics to everyday life. Asks 'what happens if' questions.
M	Very surprising that he should be so high. Literacy and numeracy poor. Fair mechanical appreciation - mechanical interests (bikes etc.).
M	Worked well in first 2 years but not working well in Yr3. Has a tendency to produce 'silly ideas' in approaches to problems. Need to keep him firmly pinned down.
M	Quiet, non-communicative, satisfactory but not special written work, little insight/imagination, conceptually very slow, not much soundly retained.
M	Doesn't like to give impression to her peer group of being interested. Easily distracted. Capable - but not special - written work. Contributes to practicals without obvious initiative. Conceptually slow

Table 7.5 is continued

Table 7.5 continued

PUPIL TYPE	DESCRIPTION
School 05 (Not a checklist school)	
FH	Better at exams than class/homework, very interested in prac. side of physics - erratic - can't be tied down to doing things, immature. Joins in with class discussion, brighter than he could be bothered to put on paper - glib conclusions - a bit arrogant.
FH	Not bad at mathematics but not good at mathematics in physics. Struggling in Yr4. Trying hard in Yr3. Lacks confidence talking to teachers - strong silent type - conforming on the whole - has gone off in Yr4 and is on report.
FH	In because of test result - not O-level material at all. Lazy, scatterbrain, son of teacher, not quite sly but nearly, not particularly good at mathematics, not dominant - a follower.
FH	Active in class discussion, ebullient, perhaps arrogant, but lazy - bad on mathematics - boarderline O-level now. Staff child.
FH	Bottom of Yr4 set, did well in 3rd Yr test, does not stand out in class - sensible - homework in on time not lazy. Shows concern to do well.
FH	Enthusiastic neat persistent questioning. Sounds less able when talked to - whiny voice, did well in 3rd Yr test - boarderline O.
M	Came for 1 year from U.S. - didn't do well in examination. As we got to know him should certainly be in top 5% [but wasn't even in R2]. Should have been picked.
M	Would not see him as top 10%. Took part in lessons but in 2nd set in Yr3 and not doing particularly well.

TABLE 7.6

SUMMARY FOR PHYSICS
Number of comments in Each Category

	SPECIFIC PHYSICS ABILITIES	GENERAL PHYSICS ABILITIES	ATTITUDE/ EFFORT/ BEHAVIOUR	OTHERS
FH	7	11	44	29
M	6	5	19	11

In the False Hit category 91 distinct comments were made spanning 20 pupils. There were:

8 (9%) references to exam performance. No pupils were picked in spite of poor exam performance.

14 (15%) comments about ability - but 6 of the 7 mentions of specific ability were to pupils being poor at something (especially at mathematics). There were 4 (4%) comments indicating a sort of 'hard slog', plodding performance.

44 (48%) comments about attitude/motivation etc. but - perhaps surprisingly these were evenly split between positive and negative. 6 pupils out of seven who were identified by teacher despite poor attitude comments were either children of staff in the school, or good on examination or both. It is interesting that attitude comments flowed freely in relation to False Hit pupils (as for Misses) but they did not seem to be of overriding importance in leading to overestimates of ability. Examination performance or parental background could override unfavourable attitudes in forming the teacher's judgement.

It is fascinating that 14 of the 20 False Hit pupils were reported by teachers to be struggling in Yr4.

In the Miss category there were 41 distinct comments referring to 13 different pupils. 7 of the comments were very individualistic (e.g. overshadowed by dominant brother and sister. Of the other comments:

19 were to do with behaviour/personality/motivation. Only 1 of these

referred to 'good' behaviour, personality, or motivation which was noticed by the teacher. Poor behaviour etc certainly seemed to be associated with pupils who were 'Misses'.

11 were related to ability: 4 referred to low specific ability but in 2 cases there were references were to high specific ability which the teacher noticed but didn't weight strongly enough to override other indications that led them to the conclusion that the pupils were not a top 10% pupil.

Intriguingly 4 of the 13 'missed' pupils were seen to be doing well in Year 4.

From the range of characteristics that teachers noticed: (attitudes, confidence, application, behaviour, neatness, seriousness, quietness) all were mentioned several times by the half dozen teachers who were interviewed, but interestingly these teachers made very little reference to things which might be expected to be more relevant to judging ability in physics - there was some reference to pupils' mathematics ability but nothing on such things as ability to handle physics concepts such as force or current, ability to suggest ways of testing an idea experimentally, or ability to interpret results in the light of experimental error - all of which have been suggested in the literature as characteristics of able science pupils and - significantly - all of which lie with the range of aims of the physics courses these teachers were teaching. Because of this they certainly could have been seen as part of the repertoire of pupils scientific behaviour that teachers might have been expected to assess - even if only in a subjective way. It is interesting that they didn't mention these things at all.

It was also interesting that when teachers talked about pupils whose ability they had overestimated, many of them mentioned characteristics like neatness, steady application, a tendency to slog on with work, and - one other that was an intriguing example of the old boy network in operation - that the pupils mother or father was a teacher or governor of the school.

However, when they were talking about pupils whose ability they had underestimated the picture was different. Here the comments normally referred to bad behaviour or to excessive quietness. These interviews suggested then that physics teachers were short of key information on their pupils and particularly aware of these seemingly less relevant characteristics. It also seemed plausible that the things which led them to underestimate ability were not just at the other end of a continuum from the things which led them to overestimate ability - but were actually of a rather different kind.

7.6 **Results for Mathematics**

TABLE 7.7

DESCRIPTIONS OF PUPILS IN MATHEMATICS BY THEIR TEACHERS

NAME	CONSTRUCTS	SPECIFIC MATHS ABILITY	GENERAL MATHS ABILITY	ATTITUDE/ BEHAVIOUR	OTHER
School 02 (Checklist)					
H	Poor attitude at times. Gen. able but ability not always easy to identify.		1	2	
H	Very keen. High gen. ability.		1	1	
FH	Works hard. Is motivated.			2	
FH	Has private tuition.		1		
FH	Keen.			1	
M	Performs to minimum expectation. Gets muddled.		2		
M	Tends to get confused.		1		
M	Not highly motivated.			1	
School 01 (Checklist)					
FH	Generally able at maths.		1		
FH	Younger than rest. Gets extra help from father.		1		1
M	Scores low on maths tests.		1		
M	Low results on maths tests.		1		
School 04 (Checklist)					
FH	Hard worker. Quiet. Does not catch on immediately.		1	2	
FH	Hard worker.			1	
FH	Good general ability.		1		
FH	Natural ability. Works hard. Needs to work slowly.		1	2	
FH	Conscientious. Not particularly inspired.		1	1	
FH	Hard worker. Younger than most in group.			1	1
FH	V.good at maths. Works hard.			1	
M	Quiet. Possible to misjudge.		1	1	

Table 7.7 is continued

NAME	CONSTRUCTS	SPECIFIC MATHS ABILITY	GENERAL MATHS ABILITY	ATTITUDE/ BEHAVIOUR	OTHER
M	Erratic work. Needs chasing. Untidy. Achieves good result when effort is put in.		2	2	
M	Sports accident meant time off school. Achieves poorer results than first year at school.		1		1
School 03 (Checklist)					
FH	Faultless effort. High achievement. Very neat.		1	2	
FH	Asks questions.			1	
M	Talks too much.			1	
M	Good exam performance.		1		
M	Does not use ability/ Poor exam performance.		1	1	
M	Big/heavy/lethargic.			1	2
School 13 (Checklist)					
H	Good potential. Rather idle. Good test performance.		2	1	
H	Hard worker. High ability.		1	1	
FH	Does well in exams. Good at arithmetic. Good in school tests.	1	2	1	
FH	Diligent. Good classwork.		1		
FH	Quite bright.				
M	Poor attitude. Not using ability.			2	
M	Poor attitude. Not using ability.			2	
H	Superficial effort. Clearly able.		1	1	
H	Poor attendance but quick with work.		1		
FH	Good at arith. Not good at anything else. Lacks confidence.	1	1	1	
FH	Above average ability. Works hard. Needs to work hard.		1	2	

Table 7.7 is continued

NAME	CONSTRUCTS	SPECIFIC MATHS ABILITY	GENERAL MATHS ABILITY	ATTITUDE/ BEHAVIOUR	OTHER
M	Low effort.			1	
M	Only superficial work. Not naturally able. Poor attitude		1	2	
M	Rushed work. Does not check work.			5	
M	Overconfident. Careless Needs self-discipline. Not enough work done.			1	
M	General improvement. Difficult to know fully - works with another pupil of similar ability.		1	1	
M	Quiet. Neat. Good at maths.		1	2	
FH	Quiet. Gets on with work. Little class-work, but good with written work.		1	3	
FH	Generally bright. Keen. Conscientious.		1	2	
FH	Good at maths. Bright. Good effort.		1	2	
FH	Good arithmetic. Under-achieving due to laziness but has potential.	1		1	
M	Good progress lately.		1		
School 06 (Not check list)					
M	Lazy. Underperforms.		1	1	
M	Quiet. Could be underperforming.		1	1	
FH	Bright. Works hard. A little underconfident.		1	2	
FH	Poor at English but good at maths.		1		1
FH	Asks if needs help. Not particularly hard-working.			2	
FH	Industrious. Not so good under pressure. Needs to build confidence.			3	
FH	Works slowly. Poor eye-sight. Methodical. Has natural insight.	1	1	1	1
FH	Difficult home environment. Not always consistent, but achieves good results quite often.		1		1

Table 7.7 is continued

NAME	CONSTRUCTS	SPECIFIC MATHS ABILITY	GENERAL MATHS ABILITY	ATTITUDE/ BEHAVIOUR	OTHER
H	Generally bright all rounder.		1		
H	Generally bright all rounder.		1		
School 00 (Not check list)					
M	Idle.			1	
M	Weak on early arith. assessment.	1			
M	Weak at maths generally.		1		
FH	Good in exams. Needs to be challenged. Can be distracted. Works hard.		1	3	
FH	Good general impression		1		
FH	Shy but above average ability.		1	1	
FH	V.neat worker.			1	
School 07 (Not check list)					
M	Poor test performance. Does not work hard enough.		1	1	
M	Quiet. Conscientious. Could be underestimated.		1	2	
M	Poor test performance.		1		
FH	Good exam performance.		1		
FH	Good test performance.		1		
FH	Good test performance in early work. Not so good now.		1		

TABLE 7.8

SUMMARY FOR MATHEMATICS
Number of comments in Each Category

	SPECIFIC MATHS ABILITIES	GENERAL MATHS ABILITIES	ATTITUDE/ BEHAVIOUR	OTHERS
FH	4	27	39	5
M	1	19	26	3
H	0	11	9	0

The first conclusion from these interviews in mathematics was that teachers tended to be very brief in their descriptions of the pupils; almost co a dismissive level. We were left with a feeling that the pupils were judged on general mathematical ability, associated, quite strongly with test performance, and general motivational characteristics. This is revealed clearly in the tables for both types of mismatch as it is also for the 'hit' category of agreement between test and teacher.

In very few cases a specific aptitude mentioned was this indicated lack of real knowledge of particular strengths or weaknesses of the pupils. Four of the five comments that we made in this category concerned arithmetical skills.

The fact that general ability was quoted as opposed to a list of specific abilities may have been as a result of the fact that many of the children who were judged as able had high ability in a large number of specific aptitudes. We doubted that the teachers could remember these specific details accurately so that the generalised comments that they made were associated in part with this. We would conjecture, therefore, that their accuracy of assessment could be high because they may notice the majority of pupils who had displayed an overall specturm of abilities, but that errors were possible because of the lack of detailed knowledge for borderline pupils.

The modification of judgement associated with motivational factors was the cause of some concern. Undoubtedly additudes to mathematics are important in that potential may not be realised if the pupil does not turn his/her attention to the subject in a positive way, yet if teachers were judging on aptitude alone, some pupils of high ability would have been missed.

Indeed co-operation in class and general willingness to apply attention to the work were at the front of teachers' minds in many cases, rather than as a secondary consideration.

7.7 General conclusions

It would be wrong to pre-empt what we might gain from a more detailed look at the things which teachers observe and use to form their judgements, and a more detailed look at the clues that the pupils give, in the classroom, to their ability, but it would be equally wrong to ignore completely what seemed to be indicated from these interviews. We should bear in mind the results from the statistical work, where we had indications of test/teacher mismatches in all subjects. This was of most concern in French and physics but of sufficient concern in all subject to warrant further identification. Our background tests covered a wide spectrum so that we could look for the biases in teachers assessment with confidence. The biases that were detected were, in many cases, not what we would expect from detailed knowledge of the pupils who were being assessed. From our interviews we would be lead to believe that this detailed knowledge was not present.

It may be that the way in which teachers were able to know their pupils was adequate for an assessment of potential for O-level to give the degree of success indicated by our statistical work, but limited to that degree. We can begin to think, more confidently, however, that this level of knowledge would build in some inaccuracies so that the test/teacher mismatch would indicate a certain amount of error on the part of teachers (in all subjects). We must not neglect test error of course, but what we are saying is that we strongly suspect teacher errors as well.

Our work should not stop with just an assessment of the errors, however, but by probing more deeply we may find further understanding that will enable more accurate identifications to be possible.

These considerations are most important when we look beyond the O-level criterion for ability, where detailed knowledge of strengths and weaknesses in the various abilities that make up a subject must be known for each pupil.

CHAPTER EIGHT

Summary of Findings from the First Year

8.1 Introduction

We have presented a large amount of evidence in previous chapters, often of a statistical kind. This is because of the care that we have taken to look quite objectively at the issues, and to leave as few stones unturned as possible. The consequence, however, is that the main ideas emerging from the findings are often lost in the detail. It is appropriate, therefore, to include two brief chapters at this point to draw together these main ideas, and then to show how we developed a research strategy for the second year of the work.

8.2 Summary

It would defeat our purpose in writing this chapter to repeat every detail of the results at this point. The reader is referred to the individual chapter summaries for a guide to the full results, and to the chapters themselves for all of the details. In this chapter we will simply reflect on the main ideas.

We recall that our research questions were set against a background of uncertainty concerning the success of teachers in identifying able children. The implications from previous research was that teachers were neither very effective, nor very efficient at identifying high IQ pupils and that alternatives to systems based on teacher judgement should be sought.

Our research differed from this earlier work in that we investigated the teachers ability to carry out identification within specific subjects, rather than their ability to identify pupils with high general ability. We can, therefore, neither confirm nor refute the earlier findings directly. However, we can conclude that, in all of the subjects we studied, the teachers' success in identifying able pupils was at a high enough level to suggest that the bleak picture painted by the earlier IQ focussed researches was not generalisable to subject specific identification. We would not wish to imply that there are no problems to be investigated, nor that teachers of these subjects can be complacent. We can say, however, that we are confident that teacher-based identification should be considered as an adequate base from which, with further understanding, one could develop a fully satisfactory method of selecting highly able pupils for whatever special provision is planned to meet their special educational needs.

We do not discount the wisdom of using test evidence to support an identification strategy based mainly on teachers' observations of their pupils. However test error does give rise to problems if a strategy is to be exclusively, or even heavily dependant on tests. Therefore we would say that our results encourage one to place confidence in the observational techniques, and suggest that these should be allowed to develop, so that experience can be gained in using teacher-based identification at least as a complement to, if not as a replacement for, test-based identification.

It is important to give this positive conclusion before discussing the shortcomings of the teacher-based method, so that the reader is left with the feeling that we are drawing conclusions about a method that needs refinement and further understanding, rather than one that should be discarded. However we now move on to some of the more detailed findings and point out some of the shortcomings of teacher-based identification that they imply.

Any prediction for future performance must have error, because of the many factors that affect individuals in different ways between prediction and final achievement. Thus test prediction (in our case the prediction for O-level score) has error. Judgements of teachers' accuracy in identifying the person who will actually achieve a high O-level score, can be made in comparison with this test prediction, but only be within the limits imposed by the accuracy of the test itself. Our attempts to take account of this point have added considerably to the complexity of earlier chapters, but they seem to us to be important and to be one way in which our work differed from much previous research. If we had asked the teachers to tell us who would be in the top band on our test scores (that is if we asked them to make some judgement of present levels of performance on very limited measures of ability) we could almost have stopped our analysis after counting the number of 'successes' in the style of Section 3.3. However our work was concerned with the more practical problem that faces a teacher at about the third form level: namely that of estimating the pupils' potential for future performance on the fairly broad canvas of a subject examination. It was in tackling this more complex, but more relevant question that we were drawn into the more detailed consideration of test error. We will, of course, continue to keep test error in mind in reporting the other findings of the study in the following paragraphs.

In all four subjects we found that high scoring pupils were missed from, and low scoring pupils were placed on, the teachers' lists. The level of mismatch between test and teacher judgement in French and physics was beyond that which could be thought of as reasonable in the light of test error. The implication, in these subjects was that teachers also made some errors (although their identification was still more successful than previous research had suggested). The mismatch was, however, 'within bounds' for mathematics and English so that, on the mismatch evidence alone, we have no cause to suspect error on the part of teachers in these subjects (though later evidence did point to the probability of such error).

Teachers, in all subjects, maintained roughly the same level of accuracy when asked to make an identification after two terms instead of one, but some changes in the lists took place. Because of the changes that occured in cases of extreme mismatch between test and teacher, we would suggest that there was a sufficient level of uncertainty in the teachers' judgements to indicate that they had made some errors in one or other of the rounds of identification in all subjects. This view of uncertainty and error can be supported by the fact that teachers were keen to make use of our test scores (we released them as soon as we had completed all work with them), and the fact that some pupils were promoted between sets as a result of test scores and then proved their worth in the new placing.

We have found very little difference between types of schools, (that is schools with a 13+ entry compared with those with an 11+ entry; large compared with small schools; or mixed compared with single sex) in the way that a teacher and an appropriate test agree on the identification of pupils of high ability. This must give particular encouragement to those in 13+ schools who often have the task of making assessments of the ability of pupils very soon after the pupils enter the school. In the case of certain subjects such as physics it could be that their task is, in effect, no different from that of third year teachers in 11+ entry schools, because it is often the case that only in the third form does this become a subject in its own right. Another reason for the similarity of success could be that even in a 11+ school the pupils change teachers from year to year so that the task of the new teacher in the identification process is the same in all situations.

We could detect some characteristics of the sort of pupils who were picked by the teachers, and these added to the picture of uncertainty connected with teachers' judgements that had emerged from the quantitative analyses. There was an implication that in all subjects the teachers displayed some individual, and in many cases, inappropriate biases in their choice of pupils. A strong leaning towards general ability in physics and French was detectable in the first round, though this tended to be less strong in the second. Pupils who had a favourable image of French were picked in that subject. These seemingly inappropriate biases on the part of teachers could provide a partial explanation for the mismatch in French and physics which was at a higher level than could be explained entirely in terms of test error. Also, as explained in Chapter Seven, they could indicate that the acceptable level of mismatch in English and mathematics may conceal some teacher error even in these subjects. Apart from the biases that were mentioned above, there was no other noticeably dominant characteristic. It is encouraging that no sex bias, no bias towards higher social class, nor to neatness or clerical speed and accuracy were detectable in individual schools (though some trends were discovered across the sample as a whole). However the most appropriate criteria for each subject were not always found to be strong characteristics of selection.

We thus suspected lack of in-depth knowledge of subject specific abilities. Support for this was given from the interviews that were held with the teachers to discuss test/teacher mismatch. General impressions of abilities and attitudes seemed to dominate the teachers perceptions of potential.

The conclusion that checklists did nothing - in the normal manner of use - to add to the success of the identification strategy was of some concern. This became less surprising when the clues that we were getting concerning the superficial knowledge of the pupils began to show. It is, however, a major conclusion for those who advocate the use of such lists and one that needs further investigation because it is unacceptable for pupils of high ability for O-level (or indeed for any pupil) to be known only superficially.

Some of our findings suggest that, until more is known about whether the insights developed in our work, and elsewhere, are helpful in fine-tuning the teacher-based system, it would be wise to ask teachers to identify a rather large target group and to make any finer selection (of say the top 2%) on the basis of further study of the nominations which they produce for this larger group. This has considerable implications for those who seek to make use of teacher nomination for small numbers of summer school places for bright pupils.

An important final point is that we are judging the 'system' as much as we are judging the teacher. Certainly we have looked at lists drawn up for a particular year group by one or more of their present teachers, but in the main our comparisons were with rank ordered test predictions on the whole of the year group. Appendix 6 contains some information to illustrate that some pupils who became test/teacher mismatches had already been put into lower teaching sets by the third year. These pupils would not have had a close scrutiny in the identification process that we initiated. Nevertheless it was the prior judgement of other teachers who put them there so we do not escape from the conclusions about teacher nomination of the most able. What we are suggesting that the process of forming the list is to some extent an evolution. This last step has the overwhelming impact of course, but we must not neglect what has led up to it.